Statistical modelling of Massively Parallel Sequencing data in forensic genetics

Vilsen, Søren B.

DOI (link to publication from Publisher):
10.5278/vbn.phd.eng.00065

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

STATISTICAL MODELLING OF MASSIVELY PARALLEL SEQUENCING DATA IN FORENSIC GENETICS

BY
SØREN B. VILSEN

DISSERTATION SUBMITTED 2018

AALBORG UNIVERSITY
DENMARK

Statistical modelling of Massively Parallel Sequencing data in forensic genetics

Ph.D. Dissertation

Søren B. Vilsen

Dissertation submitted September 23rd, 2018

Dissertation submitted:	September 23rd, 2018
PhD supervisors:	Assoc. Prof. Poul Svante Eriksen Aalborg University
	Assoc. Prof. Torben Tvedebrink Aalborg University
PhD committee:	Professor Rasmus Plenge Waagepetersen (chairman) Aalborg Universtiy
	Professor James Michael Curran University of Auckland
	Associate Professor Walther Parson Innsbruck Medical University
PhD Series:	Faculty of Engineering and Science, Aalborg University
Department:	Department of Mathematical Sciences

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-328-0

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2018

Abstract

This PhD dissertation concerns itself with the statistical modelling of DNA samples quantified by massively parallel sequencing with application to forensic genetics. The prevailing method of quantifying DNA samples found at a crime scene is capillary electrophoresis (CE). CE quantification determines the amount of DNA fragments of a given length for a set of short tandem repeat (STR) regions. However, in recent years MPS has become a viable alternative. MPS offers the ability to obtain the base compositions of the STR regions, thus, offering a higher resolution than that of CE. The aim of thesis is to lay the foundation for the statistical analysis of STR DNA mixture samples quantified by MPS.

In paper A of the thesis, we classified and modelled the systematic and non-systematic errors produced by the MPS process. The contribution of this paper is the performed analyses.

In paper B, we took a closer look at a specific type of systematic error called stutters, as they are by far the most common type of systematic error.

In paper C, we updated the DNA mixture models used for DNA samples analysed by CE to account for the results found in papers A and B. We examined the performance of the MPS model by its ability to accurately predict the probability of the amplification of alleles failing to reach a predefined (analytic) threshold.

In paper D, we constructed an evolutionary algorithm (EA) to (1) find the unknown DNA profile of a DNA sample maximising the probability of the quantitative information, and (2) approximate the probability of the evidence using only a subset of unknown DNA profile combinations.

Lastly, in paper E, we refined the MPS DNA mixture model, in-

troduced in paper C, and provide a better defined scheme for the reduction of strings exhibiting base calling errors. The data reduction scheme was an extension of a similar idea presented in paper A.

Resumé

Denne ph.d-afhandling omhandler statistisk modellering af resultater af DNA-undersøgelser af biologiske spor karakteriseret af '*massively parallel sequencing*' (MPS) med anvendelse i retsgenetik. Den fremherskende metode brugt til at karakterisere DNA fundet på et gerningssted er kapillær-elektroforese (CE). Ved CE bestemmes mængden af DNA fragmenter af en given længde for en række '*short tandem repeat*' (STR) regioner, som også kaldes mikrosatillitter. Inden for de sidste fem år er det blevet muligt at foretage DNA-sekventering ved hjælp af MPS. MPS gør det muligt at bestemme DNA-baserne i STR regionerne i stedet for blot deres længder. Derved, giver DNA-sekventering ved MPS en højere opløsning, og derfor mere information, end CE. Målet med denne afhandling er at ligge det statistiske fundament for analyse af blandinger af STR DNA fra biologiske spor med flere personers DNA, som er sekventeret med MPS.

I artikel A af denne afhandling klassificeres og modelleres de systematiske og ikke-systematiske fejl, som produceres nuder MPS-processen. Denne artikels primærer bidrag, er de udførte analyser. Ydermere kaster artikel B et nærmere blik på den hyppigste forekommende type systematisk fejl, kaldet stutter.

I artikel C modificeres DNA-mikstur-modellerne, som blev udviklet til at analysere resultater af analyser med CE. Modellen tager højde for resultaterne fra artiklerne A og B. Modellens præstationevne undersøges ved at forudsige sandsynligheden for alleler ikke bliver amplificeret over en forudbestemt beslutningsgrænse.

I artikel D udvikles en evolutionær algoritme med følgende to mål: (1) at finde den ukendte DNA-profil, som maksimerer sandsynligheden DNA-sekvensresultaterne for en DNA-prøve, og (2) at tilnærme sandsynligheden for den bevismæssigevægt ved kun at bruge en del-

mængde af de ukendte DNA-profilkombinationer.

I artikel E, modificeres MPS DNA-mikstur-modellen introduceret i artikel C, og metoden, hvormed antallet af DNA strenge som indeholder forkert kaldte baser kan reduceres. Reduktionsmetoden er en udvidelse af en koncept først skitseret i artikel A.

Contents

Thesis Details

Thesis Title:	Statistical modelling of Massively Parallel Sequencing data in forensic genetics
Ph.D. Student:	Søren B. Vilsen
Supervisors:	Assoc. Prof. Poul Svante Eriksen, Aalborg University Assoc. Prof. Torben Tvedebrink, Aalborg University

The main body of this thesis consist of the following papers.

[A] Søren B. Vilsen, Torben Tvedebrink, Helle Smidt Mogensen, and Niels Morling, "Statistical modelling of Ion PGM HID STR 10-plex MPS Data," *Forensic Science International: Genetics*, vol. 28, pp. 82–89, 2017.

[B] Søren B. Vilsen, Torben Tvedebrink, Poul Svante Eriksen, Claus Børsting, Christian Hussing, Helle Smidt Mogensen, and Niels Morling, "Stutter analysis of complex STR MPS data," *Forensic Science International: Genetics*, vol. 35, pp. 107–112, 2018.

[C] Søren B. Vilsen, Torben Tvedebrink, Poul Svante Eriksen, Claus Børsting, Christian Hussing, and Niels Morling, "Modelling allelic drop-outs in STR sequencing data generated by MPS," *Forensic Science International: Genetics*, vol. 37, pp. 6–12, 2018.

[D] Søren B. Vilsen, Torben Tvedebrink, and Poul Svante Eriksen, "DNA mixture deconvolution using an evolutionary algorithm with guided mutation," (under revision).

[E] Søren B. Vilsen, Torben Tvedebrink, Poul Svante Eriksen, Claus Børsting, Christian Hussing, and Niels Morling, "Analysis of STR DNA mixtures generated by MPS," (under revision).

In addition to the main papers, the following papers and R-packages have been published.

[1] Søren Byg Vilsen, Torben Tvedebrink, Helle Smidt Mogensen, and Niels Morling, "Modelling noise in second generation sequencing forensic genetics STR data using a one-inflated (zero-truncated) negative binomial model," *Forensic Science International: Genetics Supplementary Series*, vol. 5, pp. e416 – e417, 2015.

[2] Søren B. Vilsen, "STRMPS," found on CRAN:

https://CRAN.R-project.org/package=STRMPS.

[3] Søren B. Vilsen, "MPSMixtures," with vignette "Deconvolution of MPS STR DNA mixtures and approximation of the likelihood ratio", found on github:

https://github.com/svilsen/MPSMixtures.

This thesis was submitted for assessment in partial fulfilment o f the PhD degree. The thesis is based on the submitted or published scientific p apers, w hich a re l isted a bove. P arts o f t he p apers a re used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements were made available to the assessment committee and are also available at the Faculty.

Preface

The work presented in this thesis was carried out at the Department of Mathematical Sciences, Aalborg University, within the period September 2015 to September 2018. The PhD dissertation was supervised by associate professors Poul Svante Eriksen and Torben Tvedebrink from the Department of Mathematical Sciences, Aalborg University. Part of the research was accomplished at the School of Mathematics and Statistics at the University of Melbourne during the period August 2016 to December 2016. During the research stay, the PhD dissertation was supervised by Professor David J. Balding.

The central part of this thesis consists of a collection of five scientific papers. Three of these papers were published in peer reviewed journals, while the remaining two are undergoing revisions.

First and foremost, I would like thank my supervisors for their guidance, endless patience, wonderful discussions, and for allowing me to have an incredible amount freedom to pursue my own interests throughout the PhD period. I would also like thank Professor David J. Balding for inviting and supervising me at the University of Melbourne. Lastly, a huge thanks goes to our collaborators at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. In particular Christian Hussing, Claus Børsting, Helle Smidt Mogensen, and Professor Niels Morling, for always answering any questions, correcting manuscript aimed at cross-disciplinary journals, and for providing the data used throughout the thesis.

Søren B. Vilsen
Aalborg University, September 23, 2018

Part I

Background

Background

DNA profiling has become an essential part of the modern judicial system. If the DNA sample is of high quantity and contains DNA from a single contributor, analysis of the sample is trivial. However, DNA recovered at a crime scene is often found in low quantity, is partly degraded, contains DNA from more than one contributor or some combination thereof. When the sample contains DNA from more than one contributor, it is referred to as a DNA mixture. DNA profiles are currently mainly created by examining the length of short tandem repeat (STR) regions by capillary electrophoresis (CE). The analysis and interpretation of such profiles is a mature and on- going field of research. In recent years, massively parallel sequencing (MPS) has been introduced in forensic genetics. MPS offers the DNA base composition of the STR regions, i.e. it offers a higher resolution than that of CE. Therefore, DNA profiles obtained using MPS will offer higher discriminatory powers than those obtained with CE. It follows that developing an expert system utilizing MPS is of great interest. In order to create the foundation for such a system, we need to analyse and model the results of STR sequencing with MPS.

The remainder of Part I will consist of a short introduction to DNA, STR, polymerase chain reaction (PCR), and DNA investigation in forensic genetics, calculation of the weight-of-evidence of DNA mixture samples, evolutionary algorithms, and notes on the Poisson-gamma distribution. The first two sections were heavily based on *The Fundamentals of Forensic DNA Typing* and *Advanced Topics in Forensic DNA Typing: Methodology* by John M. Butler [6, 7].

1 DNA, STR, and PCR

An organisms deoxyribonucleic acid (DNA) contains everything necessary for passing down genetic attributes to future generations. A DNA strand can be broken down into single units of DNA, called deoxynucleotide triphosphate (dNTP), polymerised together. A dNTP consists of three parts: a nucleobase (nitrogenous bases), a deoxyribose sugar and a phosphate group as from right to left in in Fig. 1. The nucleobases will take one of four forms: adenine (A), guanine (G), cytosine (C), or thymine (T). Each of the four bases are attracted to its complimentary base, A is attracted to T and G to C, and form a base pair (bp). This attraction is what binds two DNA strands into a double helix.

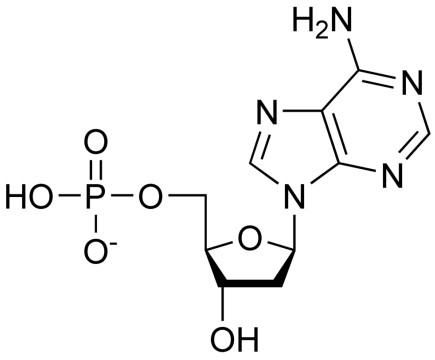

Fig. 1: A generalised version of a deoxynucleotide triphosphate. The three parts (from right to left): the nucleobase (in this case adenine), deoxyribose sugar and phosphate group. Figure stolen from https://en.wikipedia.org/wiki/Nucleotide.

The human genome contains approximately 3.2 billion bp and is organised into 22 chromosome pairs, called autosomal chromosomes, and two sex chromosomes denoted X and Y. A chromosomal location is called a marker (or a locus) and can be interpreted as a stochastic variable. The realisations of a marker are called alleles. Examining homologous chromosomes (chromosomes of the same size containing the same genetic information) will always yield exactly two alleles; these alleles can be identical or different, making the marker homozygous or heterozygous, respectively. The observed combination of alleles on a given marker is referred to as the genotype and the combination of

genotypes across multiple markers is called the DNA profile.

1.1 Short tandem repeats

STR sequences (STRs) are defined as short sequences of 2-7 bp repeated one after another (in tandem). The short, repeated sequences are called motifs. In forensic genetics, the most common motif length is four, though motifs of length three and five are also regularly employed. The following sequence is an example of an STR repeating the motif AATG six times:

$$\texttt{AATG AATG AATG AATG AATG AATG}.$$

In order to ease notation, we will condense the sequence to only its essential information and write:

$$[\texttt{AATG}]_6\,.$$

The structure of STRs can be divided into three distinct categories: Simple, compound, and complex sequences. Simple STRs contain a single motif with no inserted or missing bases, compound STRs are combinations of directly adjacent simple STRs, and complex STRs refers to the remaining sequences. In order to illustrate the differences, we give three examples:

$$\begin{aligned} \text{Simple:}\quad & [\texttt{AATG}]_6 \\ \text{Compound:}\quad & [\texttt{AATG}]_6\,[\texttt{ATTC}]_4 \\ \text{Complex:}\quad & [\texttt{AATG}]_6\,\texttt{T}\,[\texttt{ATTC}]_4\,\texttt{GGA} \end{aligned}$$

A common type of complex STRs are the so called microvariant sequences. They appear as simple or compound STRs having lost or gained one (or more) base(s). An example of a microvariant sequence is the '9.3' allele at the HUM-TH01 marker (this microvariant is present in approximately 34% of the Danish population), which has the following structure:

$$[\texttt{AATG}]_6\,\texttt{ATG}\,[\texttt{AATG}]_3$$

The sequence includes nine 'AATG' motifs and an incomplete motif of 3 bp, seen as the seventh motif missing an A. That is, the '.3' notation of the allele HUM-TH01 '9.3' refers to the number of bases in the incomplete motif.

Because of their structure, it follows that the variation between alleles of an STR marker can occur for one of the following two reasons: (1) a difference in the number of times a motif is repeated, and (2) nucleotide polymorphisms. The former results in alleles of different lengths, and the latter in alleles of the same length, but with different sequences, sometimes referred to as isoalleles.

1.2 Polymerase chain reaction

When DNA is found at a crime scene it is usually found in quantities too small to be analysed directly. Therefore, the DNA sample needs to be amplified. The polymerase chain reaction (PCR) is an in vitro method for amplifying DNA, copying marked regions of the DNA by heating and cooling the sample in a cyclic pattern, in which small stretches of DNA is double a number of times.

A region is marked using a short DNA fragment, called a DNA primer. The primer is used as both an identifier of the region and a point, from which the polymerase enzyme starts to copy the DNA fragment. Every region has two primers, a forward and a reverse primer signifying the beginning and the end of a region to be amplified. A collection of primers is called a multiplex, and may differ between manufacturers and technologies. With the current technology, CE, 10-20 regions are captured, from this point referred to as markers, while the newer technology this thesis is based on uses between 10 and 30 markers. It is worth noting that the primers are not necessarily directly adjacent to the STR region, and in such cases naturally does not just copy the STR region, but also the region between the primer and the STR region. These regions are called flanking regions.

A single cycle of the PCR process is shown in Figure 2 and can be broken into the three following stages:

(1) The sample is heated to 94 ^{o}C denaturing (separating) the two DNA strands.

(2) The sample is cooled to $50 - 60$ ^{o}C, allowing the primers to anneal (bind) to the single strands of DNA.

(3) The sample is heated to 72 ^{o}C, the DNA polymerase extends the primers copying the DNA strands.

This process is then repeated a number of times, in forensic genetics typically between 26-30 cycles. If the PCR process was 100% efficient, then 30 cycles would yield 2^{30} replicates of the sample. The PCR process efficiency is usually in the range of 0.8 to 0.9.

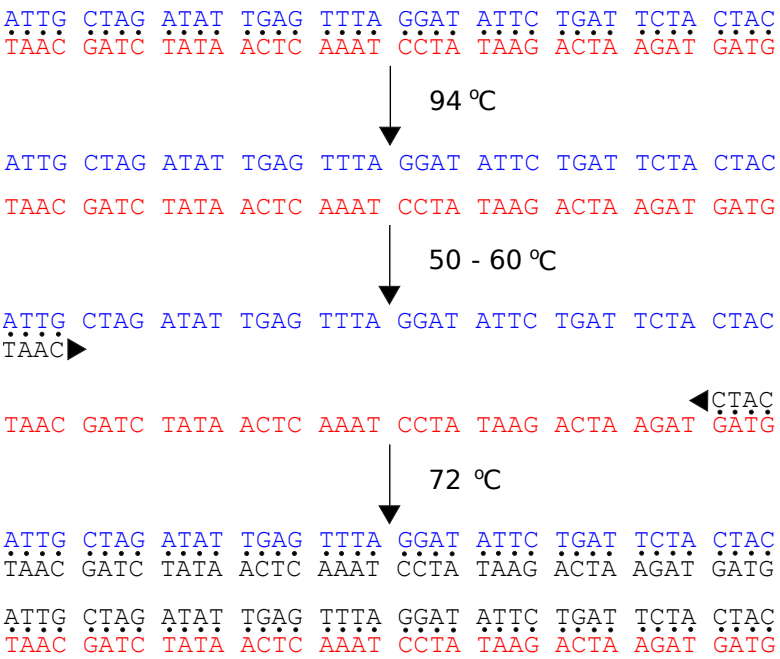

Fig. 2: The three stages of a PCR cycle: denaturing, annealing, and copying.

1.2.1 Stuttering

The most common artefact of the PCR process is stuttering. We differentiate between two types of stuttering: stuttering and back-stuttering, occurring as the loss or gain of one (or more) motif(s), respectively. The act of stuttering creates a stutter strand. We classify a stutter as single, double, triple, etc., when its has lost (or gained) 1, 2, 3, etc., motifs, respectively.

Given the DNA strand:

$$[\mathtt{AATG}]_6\,[\mathtt{ATTC}]_4\,, \tag{1}$$

an example of the possible single stutter (usually just referred to as a

stutter) of this strand is:

$$[\texttt{AATG}]_5\,[\texttt{ATTC}]_4\,,$$

i.e. the $[\texttt{AATG}]_6$ region of Eq. (1) has lost an AATG motif. A double stutter would have lost two motifs from the $[\texttt{AATG}]_6$ region, creating the strand:

$$[\texttt{AATG}]_4\,[\texttt{ATTC}]_4\,.$$

The cause of stuttering is thought to be the repetitive nature of the STR regions, and the rate with which it occurs is linked to the repetitiveness of the region, leading to the hypothesis: '*the more repetitive the region is the larger the probability of creating a stutter strand*'.

2 Analysis of the PCR product

After the DNA sample has been amplified it can be analysed. In this thesis, we exclusively analyse DNA samples investigated by massively parallel sequencing (MPS). However, before outlining the MPS process, we will give a short overview of the current state-of-the-art, capillary electrophoresis (CE).

2.1 Capillary electrophoresis

During the PCR before capillary electrophoresis (CE), primers labelled with different fluorescent chlorofores emitting light of different wavelengths are incorporated into the DNA strands. Thus, each DNA copy is characterized by its length and the fluorescent colour. The amplified DNA sample is suspended in a viscous solution and injected into a capillary electrophoresis instrument that attracts the negatively charged DNA molecules by an electric field, causing the DNA molecules to move forward in the capillary. The polymer solution in the capillary acts as sieve for the DNA strands allowing smaller DNA strands to move through the capillary faster than larger strands, and thus separating the DNA strands according to their length.

When the DNA strands have been separated, they will be detected by a CCD camera that detects the light that is emitted by the the fluorescent dyes attached to the DNA strands that are excited by a laser

The amount of light emitted will be proportional to the amount of DNA passing by the CCD camera. Thus, the signal produced by the CE process tells us the length and amount of DNA in the sample. The results are usually compared to a mixture of common alleles present in the human population for the STR markers (an allelic ladder). The results are presented as an electropherogram (Fig. 3). The amount of light representing each DNA allele (i.e. the area under the curve representing an allele) is proportional to the amount of DNA, while the height of a peak is proportional to the intensity of the DNA. When the electrophoresis is successful. The height of a peak is also proportional to the amount of DNA.

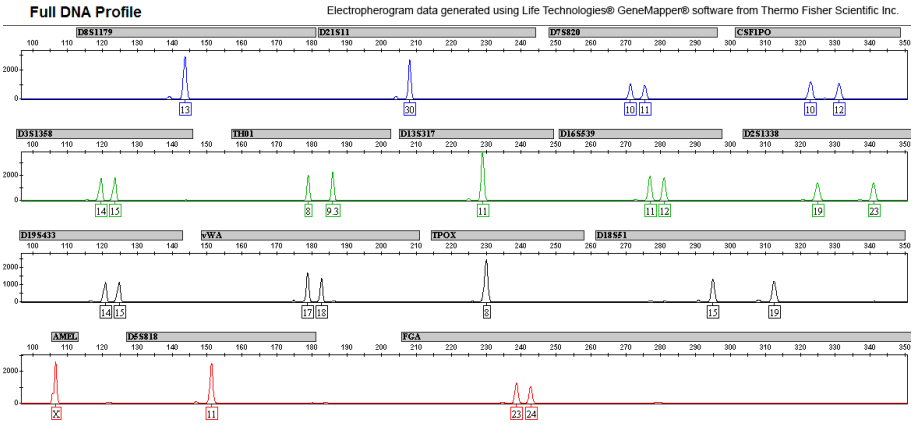

Fig. 3: An electropherogram of a DNA sample containing DNA from a single contributor. Figure stolen from `http://www.forensicsciencesimplified.org/dna/img/SInglesourceprofile.png`

2.2 Massively parallel sequencing

Massively parallel sequencing (MPS) was first introduced in 2005, as an improvement of Sanger sequencing [25].

Dependent on the manufacturer the parallelisation of the sequencing is achieved in slightly different ways, but can all be split into three stages:

(1) Library building.

(2) Clonal amplification.

(3) High throughput sequencing.

Library building relies on primers and PCR amplification, but not on fluorescent dyes. An advantage of MPS is that multiple samples can be analysed at the same time. The samples are usually distinguished by adding a *'barcode'* (different technologies utilise different nomenclature, but we will always refer to it as a barcode) to each DNA strands during the library building stage. Lastly, a key is also added, which is used for quality control. If the quality of the key-sequence is to low, then the strand is discard. In short, during library building we attach primers, barcodes, and keys to the strands and the sample is then amplified by PCR. Note because the MPS process is based on PCR, it will too suffer from stuttering.

The DNA samples analysed in this thesis were sequenced by two slightly different technologies. The first marketed by Thermo Fisher Scientific and the second by Illumina. In the first versions of commercial softwares for analysis of MPS STR data, the the sequences of primers, barcodes, and keys were removed by the software. Furthermore, the software would typically identify errors systematic and non-systematic, and try to identify the alleles of the sample under the assumption that the sample contained a single contributor. With that being said, the methodologies and implementations would differ from manufacturer to manufacturer, and only defined very loosely. However, all commercial DNA sequencing devices offers the option to export the (more or less) raw sequencing output. These are stored in a FASTQ file. A FASTQ file contains two sets of information for every strand in the DNA sample. The first corresponding to the sequenced bases of the strand and the second the quality of the sequenced bases. The two technologies are capable of sequencing DNA strands up to approximately 400 bases. Note, that the restriction on the length of the DNA strands implies that the primers used in MPS needs to be located relatively close to the STR regions. Thus, some of the primers used for CE can not be applied for MPS [8]

Systematic errors introduced during clonal amplification and sequencing can be broken into three categories:

(1) A base is called incorrectly.

(2) A base is skipped (deleted).

(3) A base in inserted.

The errors in items (2) and (3) are usually referred to as indel (short for *in*sertion and *del*etion).

In the remainder of this section, a short introduction to both of the sequencing technologies used in this thesis is given.

2.2.1 AmpliSeq - Thermo Fisher Scientific

The clonal amplification and the high throughput sequencing used in Thermo Fisher Scientific's products are emulsion PCR and ion semiconductor sequencing, respectively. The following is based on the paper by Rothberg et al. (2011) [28].

The DNA strands are amplified inside a water-in-oil emulsion PCR. Each water droplet acts as a microreactor containing the PCR reagents and ideally a single primer coated-bead with a single DNA strand. Hence, multiple PCRs can be performed simultaneously. After the emulsion breaks, the beads are covered in thousands (or millions) of copies of the original DNA strand. The beads are then placed in picoliter-volume wells containing the sequencing enzyme (DNA polymerase).

The ion semiconductor sequencing approach uses unmodified A, T, G and C – dNTPs and add them sequentially to the growing complementary DNA strand by flooding the wells. If a complementary dNTP is introduced to the next unpaired nucleotide in the original DNA strand, it is incorporated into the complementary strand by the DNA polymerase, releasing H^+ ions. In the case of homopolymer repeats, multiple nucleotides will be incorporated in a single cycle, which leads to the release of more hydrogen ions. The release of hydrogen ions is measured by a hypersensitive ion sensor, which produces an electrical signal proportional to the amount of released H^+ ions.

2.2.2 MPS with Illumina's technology

Illumina's MPS implementation involves a bridge PCR followed by sequencing-by-synthesis. The following is based on the paper by Bentley et al. (2008) [5].

During clonal amplification, an adaptor is attached to each of the DNA strands. Using these adaptors, one end of the DNA strands are

attached to a slide (also called a bridge or a flow cell). When the PCR amplification, starts the other ends of the DNA strands are attached to the slide and the strands are duplicated. The process is repeated and, thus, forms clusters of DNA strands.

In sequencing-by-synthesis, the four dNTP's are fluorescently labelled, and all are present in the reaction simultaneously. Thus, a dNTP is incorporated onto every cluster. Thereafter, the fluorescent dye is imaged in order to identify the incorporated complementary base. The dyes are then chemically removed and the process is repeated.

2.2.3 Flanking region identification

The marker and STR region of a DNA strand is identified by marker specific sequences in the flanking regions on either side of the STR region [13]. The marker specific sequences are called the forward and reverse flanking regions (flanks) of the STR region, dependent on whether they occur before or after. If both the forward and reverse flanks of a sequence (a string) are identified, we assume that the string represents that that marker. In Fig. 4 the strings s_1 and s_4 would be assumed to represent the marker, while the remaining sequences would not.

s_1 : TACACACATATGCCTA ATTG ATTG ··· ATTG GTCAGCCGGTGGAATG
s_2 : GGCCCGACTAGTGGAT CTGT CTGT ··· CTGT TGGTACTGATTGAAAC
s_3 : GATCCACATATGGCAA CTTT CTTT ··· CTTT TTTGGCCGATGGGCGA
s_4 : AATTCACATATGGTCT ATTG ATTG ··· ATTG CTTTGCCGGTGGACCG
s_5 : AGCTCTAACTTGTCAA GCTA GCTA ··· GCTA GTCTGCCGGTGGAGCC

Fig. 4: Five example strings. The red shaded area indicates the STR region. We search the strings for the forward and reverse flanks: `CACATATG` and `GCCGGTGG`, respectively. If both are found in the string, we say the string represents the marker. If a flank was found it was shaded using a light blue colour.

The distance from the end of the forward flank (and start of the reverse flank) of the STR region are known. Thus, when we have identified the marker, the STR region can easily be found, and the string can be trimmed to only include the STR region. Furthermore, when we search for reads of the forward and reverse flanks, we typically also

allow for a number of mismatches in the flanking regions. We allow for mismatches for the following two reasons: (1) the flanking regions exhibit minor variations, SNPs in the flanking regions, in the populations, and (2) the MPS process does not always determine bases correctly. In the analyses performed in this thesis, the number of allowed mismatches is always set to 1. This results in an increased estimate of '*true*' alleles in a sample, and at the '*cost*' of additional non-systematic errors.

A number of methods of analysis of MPS STR data have been offered [1, 14, 17, 22, 32–34, 36]. The author has suggested a method found in the `R`-package `STRMPS` available on `CRAN` (The Comprehensive R Archive Network: `https://cran.r-project.org/`). Of these implementations, the STRaitRazor v3 implementation is the fastest and best maintained [36]. However, STRaitRazor only exports the identified reads and completely ignores the associated quality. The quality is not necessarily needed, but it can be used to reduce the number of unique strings by identifying strings with bases called erroneously.

2.3 Major differences between MPS and CE

The major difference between the MPS and CE methods comes down to resolution. We will demonstrate some of the consequences of the added resolution by examining a sample made of 1 ng template DNA (i.e. the sample contained 1 ng of DNA fragments before PCR amplification) sequenced by the Illumina MiSeq FGx using the Illumina ForenSeq^{TM} DNA Signature Prep Kit Primer Mix A. The sample contained DNA from a single contributor. Furthermore, we will focus on the vWA marker of the sample as it exhibits all of the characteristics that we wish to highlight. In order to have a frame of reference, we will start by showing the DNA sequences aggregated by their allele length in Fig. 5. That is, a figure equivalent to what we would have obtained if the sample had been quantified by CE.

Fig. 5 shows one large bar (peak) at 17 with a coverage of 190 and what is most likely a stutter of the allele at the length 16 with a coverage of 18. This observation coupled with the fact that the sample contains DNA from a single contributor and is in large quantity, implies that the contributor is homozygous with the genotype $(17, 17)$.

If we do not aggregate the strings by allele length, we have 21

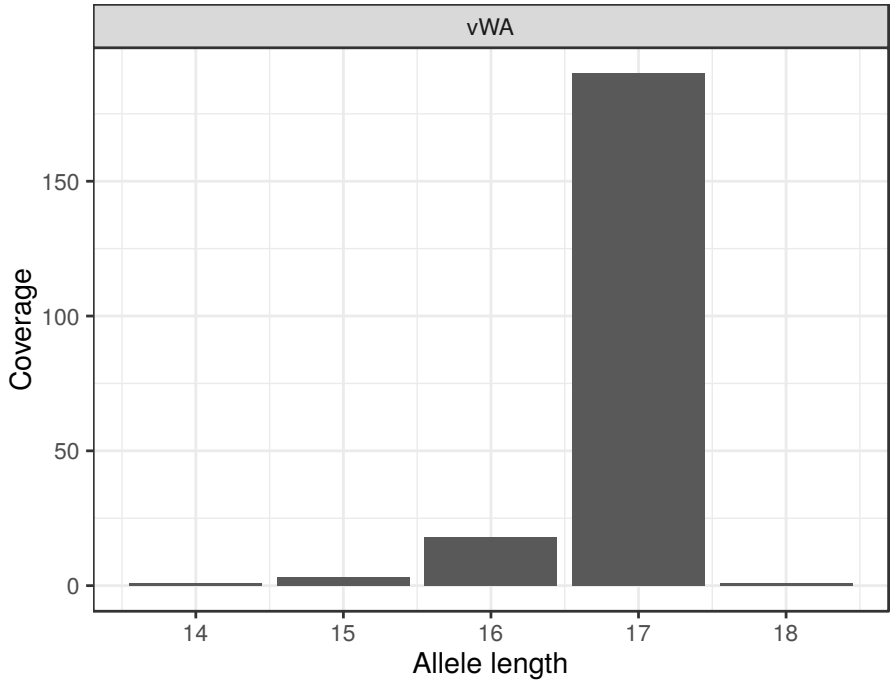

Fig. 5: The coverage against the allele length of the VWA marker of a single contributor sample investigated with 1 ng.

unique strings, and of these 15 unique strings have an allele length of 17. The two most prevalent strings are:

$$\texttt{TCTA}\ [\texttt{TCTG}]_3\ [\texttt{TCTA}]_{14}\ \texttt{TCCA}\ \texttt{TCTA} \quad \text{and} \quad \texttt{TCTA}\ [\texttt{TCTG}]_4\ [\texttt{TCTA}]_{13}\ \texttt{TCCA}\ \texttt{TCTA},$$

with coverages 95 and 80, respectively.

What looked like a homozygous marker when the data was aggregated by allele length is in fact a heterozygote marker when looking at the sequenced strings. This is sometimes referred to as isoalleles (alleles with different sequences of the same length). A few consequences of having isoalleles include:

- Larger proportion of heterozygotes in the population.
- Higher discriminatory power.
- Multiple stutter sequences.

Looking more closely at the last item, the sequenced VWA marker had two strings with allele length 16:

$$\texttt{TCTA}\,[\texttt{TCTG}]_3\,[\texttt{TCTA}]_{13}\,\texttt{TCCA TCTA} \quad \text{and} \quad \texttt{TCTA}\,[\texttt{TCTG}]_4\,[\texttt{TCTA}]_{12}\,\texttt{TCCA TCTA},$$

with coverages 10 and 8, respectively. Note that the first stutter sequence could be created by either of the two allele sequences (by loosing a TCTA motif from the $[\texttt{TCTA}]_{14}$ region or a TCTG motif from $[\texttt{TCTG}]_4$, from the first and second allele sequences, respectively), but that the second stutter sequence can only be created by the second allele sequence.

Fig. 6 is an attempt to show the unique strings of alleles with the same allele length, how they relate to the unique strings one motif shorter, and their coverage. The abscissa shows the allele length given an allele length, and each point corresponds to a unique sequence. The size of the point is proportional to its coverage on the marker. Given a sequence, A, an arrow pointing away from A to a shorter sequence, a, should be interpreted as a being a potential stutter of A. The ordinate is only used for separating the points. The points are always shown in order of prevalence with the most prevalent point closest to zero (for each length). The distance between two points depends on the coverage (relative to the entire marker) and their prevalence within strings of a given allele length (where ties are resolved randomly).

3 Weight-of-evidence for DNA mixtures

A central question when DNA evidence, $\mathcal{E}$, is presented in court is the probability of said evidence under two competing hypotheses, typically referred to as the prosecution and the defence hypotheses, denoted $\mathcal{H}_p$ and $\mathcal{H}_d$, respectively. A way of comparing the two hypotheses would be through the odds of the two hypotheses given the evidence:

$$\frac{\mathbb{P}\left(\mathcal{H}_p \mid \mathcal{E}\right)}{\mathbb{P}\left(\mathcal{H}_d \mid \mathcal{E}\right)}.$$

As it is often difficult to directly calculate the probability of an hypothesis given evidence, we apply Bayes' theorem in the numerator

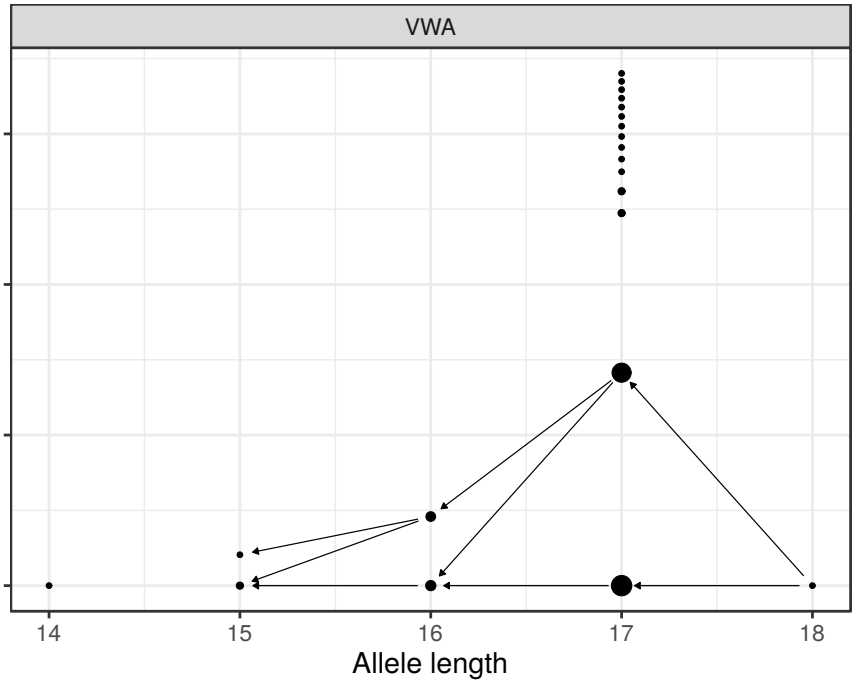

Fig. 6: A diagram of the unique strings on the vWA marker. Each point corresponds to a unique string, and its size is proportional to the relative coverage of the marker. The unique strings are ordered according to their allele lengths at the abscissa and prevalences at the ordinate (starting with the most prevalent at the bottom).

and denominator, which yields:

$$\frac{\mathbb{P}\left(\mathcal{H}_p \mid \mathcal{E}\right)}{\mathbb{P}\left(\mathcal{H}_d \mid \mathcal{E}\right)} = \frac{\mathbb{P}\left(\mathcal{E} \mid \mathcal{H}_p\right)}{\mathbb{P}\left(\mathcal{E} \mid \mathcal{H}_d\right)} \frac{\mathbb{P}\left(\mathcal{H}_p\right)}{\mathbb{P}\left(\mathcal{H}_d\right)}. \tag{2}$$

That is, we can re-write the (posterior) odds of the two hypotheses given the evidence, as the ratio of the evidence given the hypotheses, called the likelihood ratio, times the (prior) odds of the two hypotheses. We will denote the likelihood ratio of the evidence under the two hypotheses by $\mathrm{LR}(\mathcal{E}, \mathcal{H}_p, \mathcal{H}_d)$ [2, 15, 23].

Of the two quantities on the right-hand side of Eq. (2), we will focus on the likelihood ratio. In the case of DNA evidence the prior odds would represent e.g. the prior guilt of a suspect. It is not upto a statistician (or other expert witness) to interpret these prior odds. That should be left for the court to determine. Which leaves us with

the likelihood ratio.

From this point, we will assume that $\mathcal{E}$ represents quantified DNA evidence. Quantifying DNA evidence will yields some quantitative information, $\boldsymbol{y}$ (peak-heights in CE and coverage in MPS), about some combined genetic information, $\boldsymbol{g}_c$, structurally specified by a hypothesis, $\mathcal{H}_i$. Thus, given the evidence and a hypothesis, we can write the probability of the evidence given the hypothesis as:

$$\mathbb{P}\left(\mathcal{E} \mid \mathcal{H}_i\right) = \mathbb{P}\left(\boldsymbol{y}, \boldsymbol{g}_c \mid \mathcal{H}_i\right) = \mathbb{P}\left(\boldsymbol{y} \mid \boldsymbol{g}_c\right) \mathbb{P}\left(\boldsymbol{g}_c \mid \mathcal{H}_i\right), \tag{3}$$

where the last equality holds as the quantitative information is assumed to only depend on the hypothesis, $\mathcal{H}_i$, through the combined genetic information, $\boldsymbol{g}_c$.

Thus, the probability of the evidence can be factored into two parts: (1) the probability of the quantitative information given the combined genetic information, and (2) the probability of the genetic information given the hypothesis.

When the probability of the evidence is used in a DNA mixture context, the prosecution and defence may each define some set of known and unknown contributors reflecting their interpretation of the DNA mixture evidence. For simplicity, assume that we have a sample with DNA from two contributors: a victim and a perpetrator with DNA profiles $\boldsymbol{g}_v$ and $\boldsymbol{g}_p$, respectively. The police has found a suspect with DNA profile $\boldsymbol{g}_s$, who will be put on trail. Thus, the evidence in this case is $\mathcal{E} = \{\boldsymbol{y}, \boldsymbol{g}_v, \boldsymbol{g}_s\}$.

The hypotheses of the prosecution and defence are both relatively simply defined:

$\mathcal{H}_p$: The suspect is the perpetrator of the crime, $\boldsymbol{g}_s \equiv \boldsymbol{g}_p$.

$\mathcal{H}_d$: The suspect is innocent, $\boldsymbol{g}_s \not\equiv \boldsymbol{g}_p$.

Assuming that the victims DNA profile was known, we can directly calculate the probability of the evidence, under the prosecutors hypothesis, as:

$$\mathbb{P}\left(\boldsymbol{y}, \boldsymbol{g}_v, \boldsymbol{g}_s | \mathcal{H}_p\right) = \mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_v, \boldsymbol{g}_s\right) \mathbb{P}\left(\boldsymbol{g}_v, \boldsymbol{g}_s\right) \tag{4}$$

Note for convenience that we have dropped the hypothesis on the right-hand side of the equation.

We cannot do the same for the defence hypothesis. Even though the stated hypothesis seems simple, we need to properly specify what we mean by innocent. A convenient option is to assume that the perpetrator must be someone else from the population. That is, we can write the probability of the evidence under the defence hypothesis, as:

$$\mathbb{P}\left(\boldsymbol{y}, \boldsymbol{g}_v, \boldsymbol{g}_s | \mathcal{H}_d\right) = \sum_{g \in \mathcal{U}} \mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_v, \boldsymbol{g}\right) \mathbb{P}\left(\boldsymbol{g} | \boldsymbol{g}_v, \boldsymbol{g}_s\right) \mathbb{P}\left(\boldsymbol{g}_v, \boldsymbol{g}_s\right), \tag{5}$$

where $\mathcal{U}$ is the set of unknown DNA profiles in the population.

The innocence of the suspect is reflected in the first term on the right hand side, where $\boldsymbol{y}$ is independent of $\boldsymbol{g}_s$ given $\boldsymbol{g}$ and $\boldsymbol{g}_v$. From a statistical perspective, we introduce an unknown parameter, $\boldsymbol{g}$, to account for what is the unknown fragments of DNA under the defence hypothesis. Then, as we are not interested in the joint probability of the evidence and $\boldsymbol{g}$ (given the hypothesis), we remove it from consideration by using the law of total probability.

Because we will be dividing the two probabilities derived in Eq. (4) and (5) in accordance with Eq. (2), it follows that we can safely ignore the term $\mathbb{P}\left(\boldsymbol{g}_v, \boldsymbol{g}_s\right)$ found in both equations: Note that this holds as long as assumed relationship between $\boldsymbol{g}_v$ and $\boldsymbol{g}_s$ is the same under both hypotheses. It follows that, the likelihood ratio of the evidence under the two competing hypotheses simplifies to:

$$\mathrm{LR}(\mathcal{E}, \mathcal{H}_p, \mathcal{H}_d) = \frac{\mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_v, \boldsymbol{g}_s\right)}{\sum_{g \in \mathcal{U}} \mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_v, \boldsymbol{g}\right) \mathbb{P}\left(\boldsymbol{g} | \boldsymbol{g}_v, \boldsymbol{g}_s\right)}$$

This definition can be extended to any number of known and unknown DNA profiles under the two hypotheses. Assume we have a hypothesis, $\mathcal{H}_i$, specifying a set of known DNA profiles $\boldsymbol{g}_{k_i}$ and a number of unknown DNA profiles in population. If we further assume that the set of known DNA profiles under both hypotheses is denoted $\boldsymbol{g}_K = \boldsymbol{g}_{k_p}, \boldsymbol{g}_{k_d}$, we can write the general formulation of the of the probability of evidence given the hypothesis as:

$$\mathbb{P}\left(\boldsymbol{y}, \boldsymbol{g}_K | \mathcal{H}_i\right) = \sum_{g \in \mathcal{U}} \mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_{k_i}, \boldsymbol{g}\right) \mathbb{P}\left(\boldsymbol{g} | \boldsymbol{g}_K\right). \tag{6}$$

Note that the sets $\mathcal{U}$ will not only dependent on the chosen population, but also the number of unknown contributors specified by the

hypothesis. That is, if the hypothesis states that two contributors were unknown, then the set $\mathcal{U}$ is a set of 2-tuples.

We have until this point suppressed that the term $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$ will depend on unknown parameters, $\boldsymbol{\theta}$, describing the uncertainty of the measuring process. That is, Eq. (6) should be written as:

$$\mathbb{P}(\boldsymbol{y}, \boldsymbol{g}_K|\mathcal{H}_i, \boldsymbol{\theta}) = \sum_{g \in \mathcal{U}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{g}_k, \boldsymbol{g}) \, \mathbb{P}(\boldsymbol{g}|\boldsymbol{g}_K) \,. \tag{7}$$

Given the set $\mathcal{U}$, the quantified signal $\boldsymbol{y}$, and the known profiles $\boldsymbol{g}_K$, we can estimate $\boldsymbol{\theta}$ by maximising Eq. (7). That is, we choose a $\hat{\boldsymbol{\theta}}$ such that:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{g \in \mathcal{U}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{g}_k, \boldsymbol{g}) \, \mathbb{P}(\boldsymbol{g}|\boldsymbol{g}_k) \right\}.$$

Furthermore, the unknown parameters are maximised separately under each hypothesis in order to make it as fair to the suspect as possible.

Lastly, the formulas derived and presented in this section will hold for samples quantified by both CE and MPS. So where will models based on the quantified products from the two technologies differ? The answer has two parts corresponding to the two terms of the probability of the evidence given an hypothesis, seen in Eq. (3):

(1) Specifying the probability of the quantified information given the combined genetic information, $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_c)$, will need to be re-evaluated when changing from CE to MPS.

(2) The probability of the combined genetic information, $\mathbb{P}(\boldsymbol{g}_c|\mathcal{H}_i)$, will model-wise use the same formulation. Thus, we will either assume that the population is in Hardy-Weinberg equilibrium [18, 35], or account for the sub-population effects using F_{st} correction as described by Balding and Nichols [3]. The only difference is the observed allele frequencies of the population.

The change in allele frequencies of the population is entirely a consequence of the added resolution of the MPS process. That is, because we will see an increase in the number of heterozygotes and novel STR variants, when comparing MPS results with the results obtained with CE.

4 Evolutionary algorithms

When analysing a DNA sample with an unknown number of contributors, we want to be able to answer the following question: Which combination of unknown genotypes is most likely given DNA evidence? If we have the most likely combination of unknown genotypes, then we could start searching through a criminal database for matching (or partially matching) DNA profiles. This is referred to as deconvolution of DNA mixtures.

Mathematically, the question can be boiled down to finding the combination of genotypes, $\hat{g}$, which maximises the joint probability of the qualitative and quantitative results, i.e.

$$\hat{g} = \underset{g \in \mathcal{U}}{\arg\max} \left\{ \mathbb{P}\left(\boldsymbol{y}|\boldsymbol{\theta}, g_k, g\right) \mathbb{P}\left(g|g_K\right) \right\}. \tag{8}$$

Because the space of combinations of unknown genotypes, $\mathcal{U}$, is discrete, performing the maximisation in Eq. (8) would require that we examine every possible combinations of unknown genotypes. However, the size of the set, $\mathcal{U}$m is enormous, and searching through the space in this way, becomes nearly infeasible when the number of unknown contributors exceeds three. A way of solving a problem of this nature, without having to search through every possible state of the space, is to use evolutionary algorithms (EAs).

4.1 Definition of evolutionary algorithms

Evolutionary algorithms (EA) are group population based meta-heuristic optimisation methods (a branch of stochastic optimisation) [12, 21, 24, 26], which are based on the Darwinian principle of evolution: Survival of the fittest. Given a function to be maximised, called the fitness function, then an EA can be broken down as follows: (1) A population of individuals is randomly initialised, (2) the fitness of each individual is evaluated, (3) pairs of individuals are selected for breeding, these are called parents, (4) crossover is used to combine the parents into one (or more) off-spring, (5) the off-spring is mutated creating children, (6) the fitness of the children is evaluated, and (7) the new population is created by selecting individuals from the child population (sometimes the current or parent population is included in this process). Steps (2)-(7)

are then repeated until some convergence criteria has been satisfied. The pseudo code can be seen in Algorithm 1.

Algorithm 1 The General Evolutionary Algorithm.

```
1: Initialise population.
2: repeat
3:     Select parents.
4:     Cross pairs of parents.
5:     Mutate the resulting off-spring.
6:     Evaluate the fitness of the children.
7:     Select new population.
8: until Convergence.
9: return Last population.
```

From the description of the general EA, it follows that in order to implement an EA, we would need to define the following:

- **Representation**: How do we define an individual?
- **Fitness function**: What are we trying to optimise?
- **Parent selection**: How are parents selected for breeding?
- **Crossover**: How are parents combined to create off-spring?
- **Mutation**: How are the off-spring mutated?
- **Survivor selection**: How are individuals selected for the new population?

The implementation of these components will be more or less determined by the application. However, depending on the choice of individual representation, they generally fall into four overarching categories:

(1) Genetic algorithms (GAs): A string (or vector) over a finite alphabet.

(2) Evolutionary strategies (ESs): A real valued vector.

(3) Evolutionary programming (EP): Finite state machines.

(4) Genetic programming (GP): Trees.

The EA that we defined and implemented as part of this thesis to solve the problem described in the beginning of this section, is a variant of a GA. It can be shown that if the implemented genetic operators and selection methods are chosen correctly, then the GAs will converge. Convergence of GAs cannot be described in terms of rate of convergence as is the case for most stochastic optimisation methods, but a GA can be seen as a multistage Markov chain. As GAs by definition operate in a discrete state space, it follows that the Markov chain converges if, and only if, it is possible to get from one state to any other state in a finite number of steps ('state' in this context refers to an entire population) [11, 19, 29].

In terms of the operators (genetic and selection), the consequences are as follows:

- **Parent selection**: There needs to be a non-zero probability of an individual to be selected as a parent.
- **Crossover**: The probability of any combination of the two parents must exceed zero. This should include the two states where the parents are not recombined. That is, when parent 1 and 2 turn into child 1 and 2, respectively, and when parent 1 and 2 turn into child 2 and 1, respectively, without any exchange of information.
- **Mutation**: The probability of mutation is between 0 and 1 (0 and 1 not included).
- **Survivor selection**: The probability of an individual to survive to the next generation must exceed zero.

In the remainder of this section, we show a simple maximisation example using a GA with bit-string representation (known as the canonical GA) in order to show the steps necessary to implement a GA.

4.2 Example: The canonical genetic algorithm

Say, we wanted to maximise the function $f(x)$ given by:

$$f(x) = x|\sin(4x)|,$$

in some closed interval $[l; u]$.

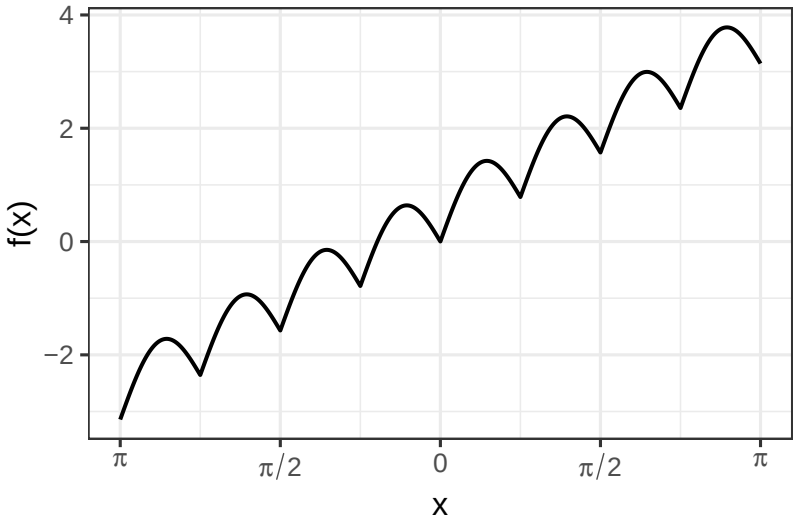

Fig. 7: The function $f(x) = x|\sin(4x)|$ shown for the interval $[-\pi; \pi]$.

The function shown in Fig. 7 exhibits two characteristics, which could cause problems for a gradient based optimisation method: (1) it is not differentiable everywhere, and (2) it contains many local maxima (sometimes referred to as the function being *'noisy'*). We could overcome the former by using a finite difference method and the latter by forcing the method to take larger steps in some intervals. However, we are still not guaranteed that a gradient based method would not get stuck in a local maximum. Therefore, we will specify a canonical GA to maximise the function.

Termination

We will terminate the algorithm after a pre-specified number of iterations.

Fitness

Defining the fitness is in this case trivial. We want to maximise $f(x)$, thus, we will use it as a fitness function.

Representation

The representation of the individual is in this case determined entirely by the choice of algorithm (i.e. canonical GA): An individual in the population is represented as a bit-string of fixed length N. The bit-string representation is a common choice as it simplifies the genetic operators (especially mutation).

It follows that an individual, $\boldsymbol{I}$, can be represented by a vector belonging to the space $\{0,1\}^N$, i.e. as:

$$\boldsymbol{I} = (i_1, i_2, ..., i_N)^T,$$

where $i_j \in \{0,1\}$.

Given the fitness function and the representation, we need a way of mapping a bit-string to an element of the interval $[l;u]$. That is, we want a map, d, satisfying:

$$d : \{0,1\}^N \mapsto [l;u].$$

The simplest solution in this instance would be to treat the bit-string as a binary encoded integer, and then rescale it to the interval $[l,u]$. Thus, we define the decoding function, d:

$$d(\boldsymbol{I}|l,u) = l + \frac{u-l}{2^N - 1} \sum_{n=1}^{N} 2^{n-1} i_n. \tag{9}$$

It should be clear from the formulation of the decoding function, d, that by choosing to use a bit-string representation, we discretise the interval $[l,u]$. Furthermore, the precision of the discretisation is completely determined by the size of the interval, $(u-l)$, and the length of the bit-string, N. That is, if we were maximising f in the interval $[-\pi;\pi]$, obtaining a precision smaller than $1/100$ would require $N \geq 10$ (if $N = 10$ the discretisation size is approximately 0.006).

Example 4.1 (Decoding an individual)
Assume that we are interested in the interval $[-\pi,\pi]$ and have proposed the candidate solution:

$$\boldsymbol{I} = (0,1,0,1,1,0,1,0,0,1)^T.$$

By Eq. (9), the decoded individual is given as:

$$\begin{aligned} x &= -\pi + \frac{2\pi}{2^{10}-1}(0+2^1+0+2^3+2^4+0+2^6+0+0+2^9) \\ &= -\pi + \frac{2\pi}{1024-1}602 \\ &\approx 0.56, \end{aligned}$$

yielding a fitness of approximately 1.35.

Crossover

The crossover operator together with the mutation operator should be designed to emulate the breeding process. In this context, the crossover operators role is to splice information from a pair of parents into two (new) individuals called the children or off-spring. Given two parents, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, we start by drawing a point called the point of crossover. The point of crossover is drawn from $\{0, 1, ..., N\}$ at random. Given the point of crossover, the parents are split at that point, and the upper half of a parent is combined with the lower half of the other parent, creating two children from two parents. A diagram of the operation is shown in Fig. 8.

$$\begin{aligned} \boldsymbol{P}_1 &= (p_{11},\ p_{12},\ \ldots,\ p_{1N})^T \\ \boldsymbol{P}_2 &= (p_{21},\ p_{22},\ \ldots,\ p_{2N})^T \\ &\downarrow \\ \boldsymbol{C}_1 &= (p_{11},\ \ldots,\ p_{1j},\ p_{2(j+1)},\ \ldots,\ p_{2N})^T \\ \boldsymbol{C}_2 &= (p_{21},\ \ldots,\ p_{2j},\ p_{1(j+1)},\ \ldots,\ p_{1N})^T \end{aligned}$$

Fig. 8: Example of crossover with the point of crossover being j of two parents of length N creating two children.

Example 4.2 (Crossover)
Assume that we are still interested in the interval $[-\pi, \pi]$. Furthermore, assume that we have two parents:

$$\boldsymbol{P}_1 = (0,1,0,1,1,0,1,0,0,1)^T \quad \text{and} \quad \boldsymbol{P}_2 = (0,0,0,0,1,1,1,0,1,0)^T.$$

Thus, the decoded parents are approximately 0.56 and -0.88, with fitness 1.35 and -0.51, respectively.
We draw randomly from the interval $[0; 10]$ the point of crossover 7 and create, as shown in Fig. 8, the two following children:

$$\boldsymbol{C}_1 = (0,1,0,1,1,0,1,0,1,0)^T \quad \text{and} \quad \boldsymbol{C}_2 = (0,0,0,0,1,1,1,0,0,1)^T.$$

When decoded, the two children yield -1.02 and 0.69 with fitness -0.22 and 1.06, respectively.

Mutation

The mutation operator's role in the breeding process is to introduce more diversity into the population. The strength of the bit-string representation is the simplicity, with which it makes the implementation of the mutation operator. Given an off-spring, $\boldsymbol{C}$, the child is mutated by running through the bit-string element-by-element and flipping a bit (changing a 0 to a 1 or vice versa) with probability $\pi^{(m)}$. For every element, c_j, we do the following:

(1) Draw a random number, $u \sim \text{Unif}(0,1)$.

(2) If $u < \pi^{(m)}$, then
$$m_j = (c_j + 1) \bmod 2.$$

Heuristic investigations have shown that the probability of flipping a bit should be around $1/N$, implying an average of 1 flipped bit per child. To the best of our knowledge, there is no theoretical result showing that this is indeed the optimal choice of $\pi^{(m)}$. Furthermore, it is likely that the optimal choice of $\pi^{(m)}$ would depend entirely on the fitness function and choice of representation.

Example 4.3 (Mutation)
Assume that we are still interested in the interval $[-\pi, \pi]$, and that we want to mutate the off-spring given by:

$$\boldsymbol{C} = (0,0,0,0,1,1,1,0,0,1)^T.$$

For each element in $\boldsymbol{C}$, we draw a random number within the unit interval and '*flip*' the entry if the random variate is less than 1 / 10. In this case, we end up flipping entries 4 and 9:

$$\boldsymbol{M} = (0, 0, 0, 1, 1, 1, 1, 0, 1, 1)^T.$$

Thus, the decoded mutated child is approximately 2.31, with a fitness of 2.49.

Selection

The way we design the selection mechanics should reflect the Darwinian principles of evolution: Survival of the fittest. That is, while the genetic operators mimic breeding, the selection mechanics should mimic natural selection. The selection mechanics, as seen above, can be split into two parts: (1) parent selection, and (2) survivor selection.

Parent selection:
The most common choice of parent selection is proportional selection (also called roulette wheel selection). In proportional selection, the probability of an individual $\boldsymbol{I}_j$ with fitness F_j being selected as a parent is:

$$\pi(\boldsymbol{I}_j) = \frac{F_j}{\sum_{k=1}^{K} F_k},$$

where K is the size of the population. That is, the higher an individual's fitness compared to the remainder of the population, the larger is its probability of being selected as a parent. It follows that the parent selection very clearly mimics natural selection. In total, we will draw K pairs of parents creating a child population of size $2K$.

Survivor selection:
At this point, we have a population of size K and a child population of size $2K$, i.e. $3K$ individuals in total. We need to reduce the size to K. These K individuals could be chosen using proportional selection, they could be chosen completely at random, or we could use an entirely different method (e.g. tournament selection, where two individuals are chosen in some way and battle for survival based on their

fitness, and the winner survives to the next population and the loser is removed from consideration).

A consequence of the randomness in both methods (in fact all three methods) is that the individual maximising the fitness function is not guaranteed survival. That is, if we find the global maximum in the descritised search space (sometimes called the 'fitness landscape'), we are not ensured that it stays in the population until convergence. A way to get around this problem is by keeping a separate super individual. The super individual represents the individual of largest fitness seen throughout the runtime of the algorithm.

We have opted for a slightly different method called elitist selection. In elitist selection, only the K individuals of largest fitness survive. This is a very strict method, which will not work well in all fitness landscapes. In fact, it does not comply with the type of survivor selection needed to have a convergent GA: The probability of survival should be non-zero. This is why using a super individual is more common. However, in this case the landscape is very simple and using elitist selection will be fine.

Results

We wanted to find the maximum of the function f in the interval $[-\pi; \pi]$. We compared the performance of the canonical GA described above with that of the optimisation method '*L-BFGS-B*' (a low memory and bounded version of the BFGS method) using finite difference to approximate the gradient. The two methods were compared using the statistical software `R` and an in-house implementations of the GA method and the `optim` function for the L-BFGS-B method.

The GA used to following settings:

- Population size: 5
- Bit-string size: 10
- $\pi^{(m)}$: 1 divided by the bit-string size.
- Number of iterations: 10

The implemented GAs and the L-BFGS-B method were randomly initialised 250 times. Note, that randomly initialising a GA implies

randomly initialising 5 individuals. Lastly, note that we implemented two versions of the GA. The first was implemented entirely in `R` [27], while the second was implemented in `C++`, `Eigen`, and `boost` [16, 30, 31] through `Rcpp`, `RcppEigen` and `BH` [4, 9, 10]. The implemented GAs can be found at:

`http://github.com/svilsen/simpleGA`.

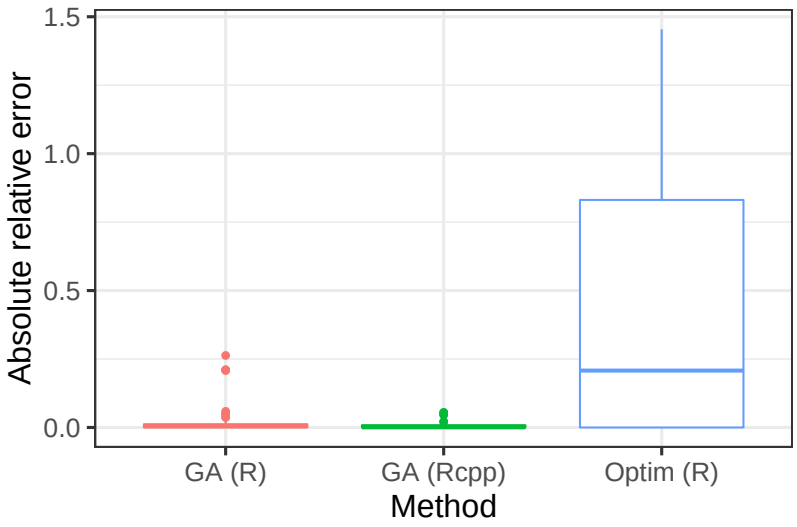

Fig. 9: Boxplots of the relative errors of the found and true maxima of f for each of the three methods using 250 simulations.

Boxplots of the absolute relative error between the maxima found by the methods and the true maximum are shown in Fig. 9. The figure shows that the medians of the two GA methods are comparable, at 0.3% and 0.5% for the `R` and `C++` implementation, respectively, (as would be expected), while the L-BFGS-B method performs poorly with a median absolute relative error of more than 20%. This is to be expected as the L-BFGS-B method is highly dependent on the starting point for noisy functions.

Fig. 10 shows boxplots of the time (in microseconds) each method took to be executed. The ordinate axis is shown on a $\log_{10}$-scale. We see that the GA (R) method was on average more than 38 times slower than the `optim` and GA (Rcpp) methods. This is again to be expected

as the `optim` and GA (Rcpp) functions are implemented in `C` and `C++`, respectively. Furthermore, we saw that the GA (Rcpp) method was slightly faster than `optim` method (1.5 times faster on average).

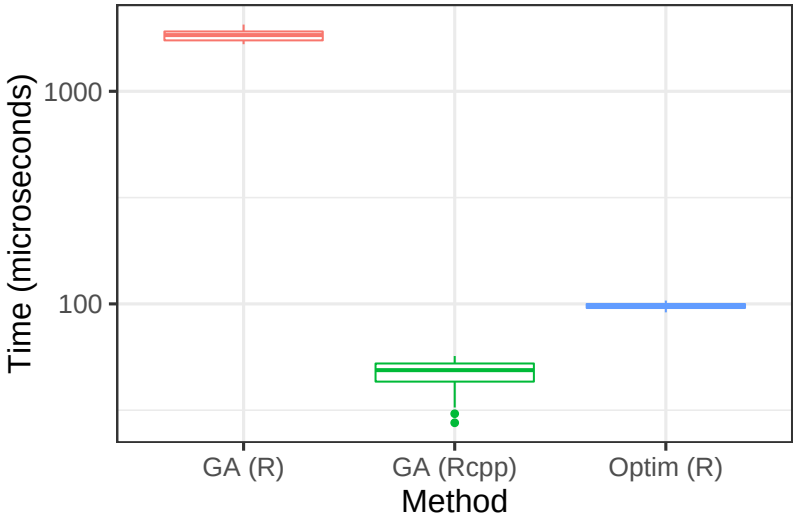

Fig. 10: Boxplots of the time it took to execute each of the three methods over 250 simulations. Simulations were made on a laptop with an Intel® Core™ i5-5300U 2.30GHz processor

5 Notes on the Poisson-gamma distribution

The quantitative results of CE investigations are continuously distributed and may be modelled according to gamma, normal, log-normal, or other relevant distributions. MPS data are expressed as counts, as the coverage of a string is just a synonym for the count of the string. It follows that the data should be modelled using a count model. The simplest choice of count model is the Poisson distribution. However, the Poisson distribution assumes that the mean and variance are equal, which is an unrealistic assumption in most situations. Therefore, this section will include notes on the Poisson-gamma distribution, that relaxes this assumption. In particular, we will derive the Poisson-gamma distribution as an overdispersed count model and introduce two variants of the Poisson-gamma distribution: (1) the Poisson-gamma distribution with a variance of order 1, and (2) the zero-truncated Poisson-

gamma distribution. This section is based on the book *'Negative Binomial Regression'* by Joseph M. Hilbe (2011) [20].

5.1 The Poisson-gamma distribution

The Poisson-gamma distribution is a mean parameterised negative binomial distribution. It is commonly interpreted as an overdispersed count model. The distribution can be derived as a hierarchical model, in which the intensity of a Poisson random variable is itself a random variable following a gamma distribution, thereby giving it its name.

Theorem 5.1
If Y is a random variable given by the following hierarchical model:

$$\begin{aligned} Y|\mu, u &\sim \text{Poisson}\,(\mu u) \\ u &\sim \Gamma\,(\theta, \theta)\,, \end{aligned}$$

where $\Gamma(\alpha, \beta)$ is the gamma distribution with shape-parameter α and rate-parameter β. Then the marginal distribution of Y is a Poisson-gamma distribution (negative binomial distribution):

$$Y \sim \text{PG}(\mu, \theta).$$

Proof. The proof follows from direct calculation of the marginal probability mass function (pmf). The marginal distribution of Y is found by integrating the joint pmf of Y and u over u:

$$\begin{aligned} p(y|\mu,\theta) &= \int_0^\infty \frac{\exp(\mu u)(\mu u)^y}{\Gamma(y+1)} \frac{\theta^\theta}{\Gamma(\theta)} u^{\theta-1} \exp(-\theta u) du \\ &= \frac{\mu^y}{\Gamma(y+1)} \frac{\theta^\theta}{\Gamma(\theta)} \int_0^\infty \exp(-(\mu+\theta)u) u^{y+\theta-1} du \\ &= \frac{\mu^y}{\Gamma(y+1)} \frac{\theta^\theta}{\Gamma(\theta)} \frac{\Gamma(y+\theta)}{(\mu+\theta)^{y+\theta}} \\ &= \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\mu}{\mu+\theta}\right)^y \left(\frac{\theta}{\mu+\theta}\right)^\theta, \end{aligned} \tag{10}$$

which is the pmf of a Poisson-gamma distribution with mean μ and overdispersion θ. □

In order to find the mean and variance of a random variable Y following in a Poisson-gamma distribution, we need the moment generating function (mgf).

Theorem 5.2
Assume Y is a random variable with pmf as defined in Eq. (10), then the mgf of Y is given as follows

$$M_Y(t) = \left(\frac{1-p}{1-p\exp(t)}\right)^{\theta}, \tag{11}$$

where $p = \mu/(\mu+\theta)$ for $t < -\log(p)$.

Proof. It follows from direct calculation of the mgf that:

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left[\exp\left(tY\right)\right] \\ &= \sum_{i=0}^{\infty} \binom{i+\theta-1}{i} p^i (1-p)^{\theta} \exp(ti) \\ &= (1-p)^{\theta} \sum_{i=0}^{\infty} \binom{i+\theta-1}{i} (\exp(t)p)^i \\ &= (1-p)^{\theta} \sum_{i=0}^{\infty} \binom{-\theta}{i} (-\exp(t)p)^i \\ &= (1-p)^{\theta} \left(1-p\exp(t)\right)^{-\theta}, \end{aligned}$$

where the second to last and last equalities holds by:

$$\binom{i+\theta-1}{i} = (-1)^i \binom{-\theta}{i},$$

and

$$\sum_{i=0}^{\infty} \binom{\theta}{i} x^i = (x+1)^{\theta},$$

respectively. □

We can now find the mean and variance of the Poisson-gamma distribution:

Theorem 5.3
Assume that Y follows a Poisson-gamma distribution with mean μ and overdispersion θ, then the mean and variance of Y are:

$$\mathbb{E}[Y] = \mu \quad \text{and} \tag{12}$$

$$\mathbb{V}\text{ar}[Y] = \mu\left(1 + \frac{\mu}{\theta}\right), \tag{13}$$

respectively.

Proof. Taking the first and second order derivative of the mgf w.r.t. t and then setting $t = 0$, we find the first and second order moments of the Poisson-gamma distribution, as:

$$M'_Y(0) = \mu \quad \text{and} \quad M''_Y(0) = \mu^2 + \frac{\mu^2}{\theta} + \mu,$$

respectively.

The mean follows directly from the first order moment, while the variance follows by inserting the first and second order moments in the definition of the variance:

$$\begin{aligned}\mathbb{V}\text{ar}[Y] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 \\ &= \mu^2 + \frac{\mu^2}{\theta} + \mu - \mu^2 \\ &= \mu + \frac{\mu^2}{\theta}.\end{aligned}$$

□

As we see from Eq. (13) the variance is dominated by the mean to the power of two. Therefore, the Poisson-gamma distribution is often called Poisson-gamma distribution of order 2 (PG2).

In order to estimate the parameters of the PG2 distribution, we utilise that the PG2 distribution can in certain situations be interpreted as belonging to the exponential dispersion family.

Theorem 5.4
If the overdispersion parameter is known, then the PG2 distribution is a member of the exponential dispersion family.

Proof. Assume that $Y \sim \text{PG2}(\mu, \theta)$. We rewrite the pmf in Eq. (10) to be in the exponential dispersion family form:

$$\begin{aligned} p(y|\mu,\theta) &= \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\mu}{\mu+\theta}\right)^y \left(\frac{\theta}{\mu+\theta}\right)^\theta \\ &= \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \exp\left\{ y \log\left(\frac{\mu}{\mu+\theta}\right) + \theta \log\left(\frac{\theta}{\mu+\theta}\right) \right\} \\ &= c(y,\theta) \exp\left\{ \eta y + A(\mu,\theta) \right\}, \end{aligned}$$

with

$$\begin{aligned} \eta &= \log\left(\frac{\mu}{\mu+\theta}\right), \\ A(\mu,\theta) &= -\theta(1-\exp(\eta)), \\ c(y,\theta) &= \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}, \quad \text{and} \\ \phi &= 1. \end{aligned}$$

If the overdispersion parameter, θ, is known, then the PG2 distribution is a member of the exponential dispersion family. □

Because the PG2 distribution is an the exponential dispersion family when θ is known, the parameters of the PG2 distribution are typically estimated by assuming $\mu = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$ and alternating between the two following states:

(1) Given the current estimate of θ: Estimate $\boldsymbol{\beta}$ by iteratively reweighted least squares (IRLS).

(2) Given the current estimate of $\boldsymbol{\beta}$: Estimate θ by setting the derivative of Eq. (10) w.r.t. θ equal to 0 and use the Newton-Raphson method to find the θ maximising Eq. (10).

A popular alternative to item (2) is to use deviance based estimation equivalent to how the dispersion parameter is estimated in a quasi-Poisson models.

5.2 The Poisson-gamma distribution of order 1

A downside to the PG2 distribution is that if the mean is large, then the variance is very large. Therefore, we define the Poisson-gamma

distribution of order 1 (PG1). The PG1 distribution can be defined in multiple ways, but the simplest definition is in terms of the PG2 distribution.

Definition 5.1
We say that Y follows a PG1 distribution with parameters μ and γ if Y follows a PG2 distribution with parameters μ and $\theta = \mu/\gamma$.

That is, the PG1 distribution is interpreted as a PG2 distribution, where the overdispersion parameter is dependent on the mean. As with the PG2 distribution, the PG1 distribution gets its name by because the variance of a random variable following the PG1 distribution will be dominated by the mean to the power of 1.

Theorem 5.5
If $Y \sim \text{PG1}(\mu, \gamma)$, then the pmf is given as:

$$p(y|\mu,\gamma) = \frac{\Gamma\left(y + \frac{\mu}{\gamma}\right)}{\Gamma\left(y+1\right)\Gamma\left(\frac{\mu}{\gamma}\right)} \left(\frac{1}{\gamma+1}\right)^{y+\mu/\gamma} \tag{14}$$

and the mean and variance of Y are:

$$\begin{aligned} \mathbb{E}\left[Y\right] &= \mu \quad \text{and} \\ \mathbb{V}\text{ar}\left[Y\right] &= \mu\left(1+\gamma\right), \end{aligned}$$

respectively.

Proof. The proof is similar to the proof of Theorem 5.3. □

Even though this formulation of the Poisson-gamma distribution is very useful, it is worth noting the following: The PG1 distribution can never be a member of the exponential dispersion family, because the overdispersion depends directly on the mean. Therefore, the parameter estimation cannot be split into two parts, contrary to the PG2 distribution, but it has to be performed simultaneously. That is, parameter estimation is more difficult, and the results are less reliable.

5.3 The zero-truncated Poisson-gamma distribution

The zero-truncated Poisson-gamma distribution (ZTPG) is useful when modelling the coverage of strings, which cannot be classified as alleles and are most likely not systematic errors (primarily stutters). These needs to be zero truncated as the outcome zero is not possible in this context. As with the PG1 distribution, the pmf of the ZTPG is found in terms of the PG2 distribution.

Theorem 5.6
If Y follows a zero-truncated Poisson-gamma distribution, with mean ω and overdispersion α, then the pmf is given by:

$$p(y|Y>0,\omega,\alpha) = \frac{\dfrac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\left(\dfrac{\omega}{\omega+\alpha}\right)^{y}\left(\dfrac{\alpha}{\omega+\alpha}\right)^{\alpha}}{1-\left(\dfrac{\alpha}{\omega+\alpha}\right)^{\alpha}}. \tag{15}$$

Proof. The pmf of any zero-truncated discrete distribution can be written as

$$\mathbb{P}\left(Y=y|Y>0,\boldsymbol{\theta}\right) = \frac{\mathbb{P}\left(Y=y|\boldsymbol{\theta}\right)}{\mathbb{P}\left(Y>y|\boldsymbol{\theta}\right)} = \frac{\mathbb{P}\left(Y=y|\boldsymbol{\theta}\right)}{1-\mathbb{P}\left(Y=0|\boldsymbol{\theta}\right)}.$$

Eq. (15) is obtained by inserting the pmf of the PG2 distribution in place of $\mathbb{P}\left(Y=y|\boldsymbol{\theta}\right)$ and $\mathbb{P}\left(Y=0|\boldsymbol{\theta}\right)$, with $\boldsymbol{\theta}=(\omega,\alpha)^T$. □

As with the PG2 distribution, we need the mgf of the ZTPG to find the mean and variance.

Theorem 5.7
Assume that Y follows a ZTPG distribution with mean ω and overdispersion α, then the mgf of Y is given as follows

$$M_Y(t) = \frac{(1-p)^{\alpha}}{1-(1-p)^{\alpha}}\Big[\left(1-p\exp(t)\right)^{-\alpha}-1\Big], \tag{16}$$

where $p=\omega/(\omega+\alpha)$ for $t<-\log(p)$.

Proof. The proof is similar to the proof of Theorem 5.2. □

From the mgf we can find the mean and variance of the ZTPG.

Theorem 5.8
If Y follows a ZTPG distribution with mean ω and overdispersion α, then its mean and variance are given by

$$\mathbb{E}[Y|Y>0] = \frac{\omega}{(1-(1-p)^{\alpha})}, \quad \text{and}$$
$$\mathbb{V}\text{ar}[Y|Y>0] = \frac{p\alpha - (p\alpha)^2(1-p)^{\alpha} - p\alpha(1-p)^{\alpha}}{(1-p)^2(1-(1-p)^{\alpha})^2},$$

respectively, where $p = \omega/(\omega+\alpha)$.

Proof. The proof is similar to that of Theorem 5.3 with the first and second order moments given by:

$$M'_{Y|Y>0}(0) = \frac{p\alpha}{(1-p)(1-(1-p)^{\alpha})}, \quad \text{and} \tag{17}$$
$$M''_{Y|Y>0}(0) = \frac{p\alpha(p\alpha+1)}{(1-p)^2(1-(1-p)^{\alpha})}. \tag{18}$$

□

Parameter estimation for the ZTPG distribution is, as with the PG1 distribution, difficult. The parameters could be estimated by maximising the parameters simultaneously, using a sophisticated quasi-Newton method. However, a viable alternative is to use iterative method-of-moments based estimation:
Assuming $y_n \sim \text{ZTPG}(\omega, \alpha)$ for $n = 1, ..., N$, the moment estimates of ω and α can be found be iteratively solving the equations:

$$0 = \left(\frac{1}{N}\sum_n y_n\right)(1-(1-p)^{\alpha}) - \omega, \text{ and} \tag{19}$$
$$0 = \left(\frac{1}{N}\sum_n y_n^2\right) - \left(\frac{1}{N}\sum_n y_n\right)^2 - \frac{p\alpha - (p\alpha)^2(1-p)^{\alpha} - p\alpha(1-p)^{\alpha}}{(1-p)^2(1-(1-p)^{\alpha})^2}, \tag{20}$$

for ω and α, respectively.

The parameters that solve these equations can be found using the Newton-Raphson method. This implies that we need the derivatives of the two equations. The derivative of Eq. (19) w.r.t. ω is:

$$\left(\frac{1}{N}\sum_n y_n\right)(1-p)^{\alpha+1} - 1$$

The derivative of Eq. (20) w.r.t. α is very complicated. The most simplified form is:

$$\_\frac{p^2\omega(\alpha^2(s+1)s+\alpha vs+v^2)+p^2\alpha vu+p\alpha(\alpha v+\omega(u+\alpha))s\log(1-p))}{\alpha^2(\omega+\alpha)^3v^3}$$

where $s=(1-p)^\alpha$, $v=s-1$, and $u=\alpha s+v$.

6 Organisation of the remainder of the thesis

The aim of this section is to outline the relationship between the information provided in Part I with the five papers included in Part II of the thesis.

- **Paper A**: The aim of the paper was to classify and model the systematic and non-systematic errors produced by the MPS process, presented in Sections 1.2.1 and 2.2. Furthermore, the behaviour of the quality, mentioned in Section 2.2, was also analysed. The primary contribution of this paper are the performed analyses. While the analyses are simple, they are extremely useful for the statistical modelling presented in later papers.

- **Paper B**: The main objective of the paper was to construct a predictor of stuttering, introduced in Section 1.2.1, which utilised the added resolution of the MPS process. The reason for this construction is two fold: (1) The added resolution can be used to give more weight to the hypothesis '*the more repetitive the region is the larger the probability of creating a stutter strand*', and (2) Stutters are by far the most common type of systematic error, therefore, understanding stutters in the MPS setting is integral to the modelling of samples quantified by MPS.

- **Paper C**: We update the DNA mixture models used for DNA samples quantified by CE to account for the results found in papers A and B. We examine the performance of the MPS model by its ability to accurately predict the probability of the amplification of alleles failing to reach a predefined (analytic) threshold. The accuracy is assessed by comparing the observed Brier scores to the expected Brier scores under the assumption that the model is accurate. The contributions of the paper are the MPS DNA mixture model and the derivation of the mean and variance the Brier score.

- **Paper D**: We constructed an evolutionary algorithm (EA) to solve the objectives outlined at the end of Section 3 and the beginning of Section 4. That is, to (1) find the unknown DNA profile (or combination of unknown DNA profiles) of a DNA sample maximising the probability of the quantitative information, called deconvolution, and (2) approximate the probability of the evidence using only a subset of the (set of) unknown DNA profile combinations. Furthermore, we show that we can obtain quicker and more reliable deconvolution, by using residuals of the model to guide the genetic operators of the EA.

- **Paper E**: We took a critical look at the MPS DNA mixture model, introduced in paper C, and refined it by changing the distribution of the quantitative information and level of stutter recursion included in the model. Furthermore, we provide a better defined scheme for the reduction of strings exhibiting base calling errors. A scheme loosely presented in paper A.

Lastly, it should be noted that the workflow used analyse the all data in these papers, was implemented in `R` as the package `STRMPS` which is available on `CRAN` (The Comprehensive R Archive Network: `https://cran.r-project.org/`). Furthermore, the evolutionary algorithm used in paper D was primarily implemented in `C++` with an `R` interface through `Rcpp`, as the package `MPSMixtures`, available at `https://github.com/svilsen/MPSMixtures`.

References

[1] S. Y. Anvar, K. J. van der Gaag, J. W. F. van der Heijden, M. H. A. M. Veltrop, R. H. A. M. Vossen, R. H. de Leeuw, C. Breukel, H. P. J. Buermans, J. S. Verbeek, P. de Knijff, J. T. den Dunnen, and J. F. J. Laros, "Tssv: a tool for characterization of complex allelic variants in pure and mixed genomes," *Bioinformatics*, vol. 30, no. 12, p. 1651, 2014. [Online]. Available: +http://dx.doi.org/10.1093/bioinformatics/btu068

[2] D. J. Balding, *The Weight-of-evidence for Forensic DNA profiles*. Wiley, 2005.

[3] D. Balding and R. Nichols, "DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands," *Forensic Science International*, vol. 64, pp. 125 – 140, 1994.

[4] D. Bates and D. Eddelbuettel, "Fast and elegant numerical linear algebra using the RcppEigen package," *Journal of Statistical Software*, vol. 52, no. 5, pp. 1–24, 2013. [Online]. Available: http://www.jstatsoft.org/v52/i05/

[5] D. R. Bentley *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, pp. 53 – 59, 2008. [Online]. Available: http://dx.doi.org/10.1038/nature07517

[6] J. Butler, *Fundamentals of Forensic DNA Typing*. Academic Press, 2009.

[7] ——, *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, 2012.

[8] C. Børsting and N. Morling, "Next generation sequencing and its applications in forensic genetics," *Forensic Science International: Genetics*, vol. 18, pp. 78 – 89, 2015.

[9] D. Eddelbuettel, J. W. Emerson, and M. J. Kane, *BH: Boost C++ Header Files*, 2018, R package version 1.66.0-1. [Online]. Available: https://CRAN.R-project.org/package=BH

[10] D. Eddelbuettel and R. François, "Rcpp: Seamless R and C++ integration," *Journal of Statistical Software*, vol. 40, no. 8, pp. 1–18, 2011. [Online]. Available: http://www.jstatsoft.org/v40/i08/

[11] A. E. Eiben, E. H. L. Aarts, and K. M. Van Hee, "Global convergence of genetic algorithms: A markov chain analysis," in *Parallel Problem Solving from Nature*, H.-P. Schwefel and R. Männer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991, pp. 3–12.

[12] A. Eiben and J. Smith, *Introduction to Evolutionary Computing*. Springer, 2003.

[13] S. Fordyce, M. Avila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F. Petersen, E. Willerslev, A. Hansen, N. Morling, and M. Gilbert, "High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform," *BioTechniques*, vol. 51, pp. 127 – 133, 2011.

[14] S. L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, and N. Morling, "Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs," *Forensic Science International: Genetics*, vol. 21, pp. 68–75, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2015.12.006

[15] I. J. Good, "Probability and the weighing of evidence," *Philosophy*, vol. 26, no. 97, pp. 163–164, 1951.

[16] G. Guennebaud, B. Jacob *et al.*, "Eigen v3," http://eigen.tuxfamily.org, 2010.

[17] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich, "lobSTR: A short tandem repeat profiler for personal genomes," *Genome Research*, vol. 22, no. 6, pp. 1154–1162, 2012.

[18] G. Hardy, "Mendelian Proportions in a Mixed Population," *Science*, vol. 28, pp. 49 – 50, 1908.

[19] J. He and L. Kang, "On the convergence of genetic algorithms," *Theoretical Computer Science*, vol. 229, pp. 23–29, 1999.

[20] J. M. Hilbe, *Negative Binomial Regression*, 2nd ed. Cambridge University Press, 2011.

[21] J. H. Holland, *Adaptation in Natural and Artificial Systems*. MIT Press, 1975.

[22] J. L. King, F. R. Wendt, J. Sun, and B. Budowle, "STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems," *Forensic Science International: Genetics*, vol. 29, pp. 21–28, 2017.

[23] D. V. Lindley, "A problem in forensic science," *Biometrika*, vol. 64, no. 2, pp. 207–213, 1977. [Online]. Available: http://www.jstor.org/stable/2335686

[24] S. Luke, *Essentials of Metaheuristics*, 2nd ed. Lulu, 2013, available for free at http://cs.gmu.edu/∼sean/book/metaheuristics/.

[25] M. Margulies *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, p. 376, 2005.

[26] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1996.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[28] J. M. Rothberg *et al.*, "An integrated semiconductor device enabling non-optical genome sequencing," *Nature*, vol. 475, pp. 348 – 352, 2011. [Online]. Available: http://dx.doi.org/10.1038/nature10242

[29] G. Rudolph, "Convergence Analysis of Canonical Genetic Algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 1, pp. 96–101, 1994.

[30] B. Schling, *The Boost C++ Libraries*. XML Press, 2011.

[31] B. Stroustrup, *The C++ Programming Language*, 4th ed. Addison-Wesley Professional, 2013.

[32] C. Van Neste, M. Vandewoestyne, W. Van Criekinge, D. Deforce, and F. Van Nieuwerburgh, "My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing," *Forensic Science International: Genetics*, vol. 9, no. 1, pp. 1–8, 2014.

[33] D. H. Warshauer, J. L. King, and B. Budowle, "STRait Razor v2.0: The improved STR Allele Identification Tool-Razor," *Forensic Science International: Genetics*, vol. 14, pp. 182–186, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2014.10.011

[34] D. H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, J. L. King, and B. Budowle, "STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data," *Forensic Science International: Genetics*, vol. 7, no. 4, pp. 409–417, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2013.04.005

[35] W. Weinberg, "On the Demonstration of Heredity in Man," 1908, Boyer SH, ed., *Papers on human genetics*. Englewood Cliffs, NJ: Prentice Hall. Translated, 1963.

[36] A. E. Woerner, J. L. King, and B. Budowle, "Fast STR allele identification with STRait Razor 3.0," *Forensic Science International: Genetics*, vol. 30, pp. 18–23, 2017.

Part II

Papers

Paper A

Statistical modelling of Ion PGM HID STR 10-plex MPS Data

Søren B. Vilsen, Torben Tvedebrink, Helle Smidt Mogensen, and Niels Morling

The paper has been published in the
Forensic Science International: Genetics, vol. 28, pp. 82–89, 2017.

The layout has been revised.

Abstract

We investigated the results of short tandem repeat (STR) markers of dilution series experiments and reference profiles generated using the Ion PGM massively parallel sequencing platform utilising the HID STR 10-plex panel. The STR markers were identified by the marker specific flanking regions of the STR region.

We investigated the following: (1) The usage of quality measures for identifying substitution errors, (2) the heterozygote balance and compared it to that of capillary electrophoresis (CE), (3) the stability of the coverage and the consequence of IonExpress Barcode adapter (IBA) sampling with decreasing amounts of template DNA, (4) the hypothesis that the parental longest uninterrupted stretch (LUS) is a better linear predictor of stutter ratio than the parent allele length, (5) the use of parental allele length as a predictor of shoulder ratio, and (6) the removal of non-systematic erroneous sequences using dynamic thresholds created by fitting the distribution of the non-systematic erroneous sequences.

We found that, due to MID sampling, the average coverage on a marker could not be used as an apt predictor of the amount of template DNA. The parental LUS was shown to be better predictor of stutter ratio than the parental allele repeat length, when markers with compound and complex repeat patterns or markers which contained micro-variants were considered, such as marker TH01 showed R^2 of 0.02 and 0.78 for parent allele repeat length and LUS, respectively. The one-inflated negative binomial method (OINB) and geometric model that can be used to remove non-systematic noise left on average 1.8 and 1.2 systematic errors per STR system, respectively.

1 Introduction

In forensic genetics, the STR regions are typically analysed by examining the length of the regions by capillary electrophoresis (CE) [1, 2]. In recent years, massively parallel sequencing (MPS), also known as Next- or Second-Generation Sequencing, has been introduced in forensic genetics [3–16]. MPS offers the base composition of the STR regions, i.e. it offers a higher resolution than that of CE. Therefore, DNA profiles obtained using MPS will offer higher discriminatory powers. It follows that developing an expert system utilizing MPS is of great interest. In order to create such a system, we need to analyse and model the results of STR sequencing with MPS.

As sequencing STRs with MPS is still in its early stages, most of the research has not focused on modelling the artefacts. The majority of the research has been focused on (1) showing that STRs could be sequenced using MPS, (2) analysing the observed STR sequence variations [3–5, 9–12], and (3) creating methods for easily identifying STR variants and sample genotyping [6–8, 13, 14, 17]. With the introduction of the first commercial MPS STR typing kit for forensic genetics, the Ion Torrent^{TM} HID STR 10-plex from Thermo Fisher, we will soon be able to add a fourth category: Evaluating STR MPS typing kit [16, 17].

The aim of this article is to establish the foundation for STR MPS expert systems. With this in mind, we are going to investigate the data presented in Fordyce et al. (2015) and Friis et al. (2016) by analysing the following:

(1) The effect of marker and STR region identification on the quality of reads.

(2) The influence of the average allele coverage on the variability of the heterozygote balance.

(3) The influence of the amount of DNA on the coverage.

(4) The coverage of stutters and shoulders compared to those of the parental alleles.

(5) The coverage distribution of the non-systematic erroneous sequences.

The general structure of the paper is as follows: The experiments and the generated data are presented in Sections 2.1 and 2.2. Section

2.3 introduces the statistical methods used to analyse the data. Results of the performed analyses are found in Section 3. A discussion is found in Section 4.

2 Materials and methods

2.1 Experiments

The experimental data consisted of two parts: Dilution series (2.1.1) and reference samples (2.1.2). If having a large number of different alleles was important for the focus of the analyses, e.g. stutter ratios, the reference samples were used. When assessing the variability of a quantity across a range of DNA concentrations, e.g. risk of allelic drop-out, the dilution series data were used. The experimental data used in this manuscript have previously been analysed by Fordyce et al. (2015) [16] and Friis et al. (2016) [17]. Furthermore, unless otherwise explicitly stated, we used the experimental data.

2.1.1 Dilution series experiments

The dilution series experiments ranged from 50pg to 2ng DNA, containing six dilutions approximately halving the amount of DNA each time, or more precisely each series contained samples at 2ng, 1ng, 500pg, 200pg, 100pg, and 50pg DNA. Four series were created from two contributors in two experiments. The DNA strands were sequenced with the Ion Torrent^{TM} HID STR 10-plex and the Ion PGM^{TM} (Thermo Fisher Scientific) for all four experiments with two experiments per Ion 318v2 chip, one from each contributor using IonExpress Barcode Adaptors (IBA) to differentiate between contributors and dilutions.

2.1.2 Reference data

The reference files included results from DNA samples (randomly) obtained from 207 individuals from the Danish population. In total, 13 pooled libraries were created each containing 16 samples. These libraries were sequenced on 318v2 chips using the Ion Torrent^{TM} HID STR 10-plex panel and the Ion PGM^{TM} [17].

2.1.2.1. Ethical considerations

The work was approved by the Danish ethical committee (H-1-2011-081).

2.2 Data

The data were provided in a raw FASTQ [18] format with unaligned sequences and corresponding quality assignments. Preprocessing was performed by the Torrent Suite™ software that filtered and trimmed the sequences in accordance with the default settings of the software. In particular, the software removed (1) polyclonal beads, (2) wells with phasing, (3) sequences of general low quality, and (4) trimmed the 3' end to an acceptable level of quality. For a more detailed description see the Torrent Suite™ technical notes [19].

During the MPS workflow, stutters are created. Typically, stutters occur as sequences four bases shorter or longer (assuming tetranucleotide markers) than the parental allele [20].

Artefacts of the MPS workflow are insertion, deletion, and substitutions. Substitutions result in sequence variations, i.e. sequences of lengths equal to those of the true alleles, but with one or more base differences in the sequences. The substitution of a base should ideally be reflected in the quality of the sequence. The consequence of an insertion or deletion is a sequence one or more bases longer or shorter, respectively, than the analysed strand. We refer to these as shoulders (right and left shoulders of insertions and deletions, respectively).

2.3 Statistical methods

2.3.1 Marker and STR region identification

Our analysis was based on unaligned FASTQ files. Therefore, we needed to identify the STR regions. This was achieved by identifying the STR region by searching for marker specific forward and reverse sequences in the flanking regions adjacent to the STR regions [3]. A sequence was identified if the following four conditions were met:

(1) The forward flank was observed before the reverse flank.

(2) The forward and reverse flanks was observed exactly once.

(3) Only the flanks of a single marker was found.

(4) No more than one mismatch was observed in the flanking regions.

The sequences corresponding to the reverse complementary strand were analysed in the same manner as described above by reversing the forward/reverse flanking regions and finding their complementary.

2.3.2 The quality

When a base is called by the Ion Torrent software, a quality score is also reported. The quality score is usually defined in one of two different ways, both of which are based on the estimated probability of error. The Ion Torrent software uses the Phred score [21]:

$$Q = -10 \log_{10} \mathbb{P}(\text{Error}). \tag{A.1}$$

The probability of error, $\mathbb{P}(\text{Error})$ or $\mathcal{P}$ for short, was estimated based on six predictors of local quality and a lookup table [22]. Intuitively, the quality score should decrease when a base is called erroneously, however, this is not always observed for miscalled bases. The quality score tends to decrease with each sequenced base as the risk of error accumulates with the number of bases sequenced.

Our main interest in the quality lies in its use in identifying base calling errors. If we knew whether a base calling error had occurred, we could augment the coverage of the true strand with the coverage of the erroneous strands or simply remove the erroneous strands from consideration. However, this information is unavailable. Using the quality score, $\mathcal{Q}$ (or equivalently $\mathcal{P}$), we can assign this a probability.

The simplest approach would be to examine the effect of base calling errors on the quality of the entire sequence. We defined the quality of an entire sequence as the geometric mean of the base qualities of the sequence. We used the geometric mean as it is more sensitive to outliers than the arithmetic mean. Bases with low quality scores will, therefore, have a larger impact on the quality of the sequence. Thus, given a sequence, $\mathcal{S}_i$, of length $|\mathcal{S}_i|$, its quality is:

$$\mathcal{Q}(\mathcal{S}_i) = \left(\prod_{j=1}^{|S_i|} q_{ij} \right)^{1/|S_i|}, \tag{A.2}$$

where q_{ij} is the quality of the j'th base of the i'th sequence.

We would like to assess if the quality drops at or around a base substitution. We did so by comparing the base quality of the most prevalent sequence with those of all other unique sequences of given markers and of specific lengths. Therefore, we needed the aggregate base qualities of every unique sequence and every base in those sequences. As with the sequence quality, we used the geometric mean of the base qualities. Given a set of sequences of equal length, $\mathcal{S}_\mathcal{I} = \{\mathcal{S}_i\}_{i\in\mathcal{I}}$ for some set of indices $\mathcal{I}$, we defined the j'th base quality, $\mathcal{B}_j$, of that set in a similar way as that in Eq. (A.2):

$$\mathcal{Q}(\mathcal{B}_j; \mathcal{S}_\mathcal{I}) = \left(\prod_{i\in\mathcal{I}} q_{ij}\right)^{1/|\mathcal{I}|}. \tag{A.3}$$

This is also called the second dimension of quality. If the sequence set used was clear from the context, the base quality was denoted $\mathcal{Q}(\mathcal{B}_j)$. Note that if the sequences were ordered into a matrix, the $\mathcal{Q}(\mathcal{S}_i)$ would be the row average and $\mathcal{Q}(\mathcal{B}_j)$ the column average.

In order to examine the difference in base quality between two sets of sequences (assuming the sequences are of equal length), we define the quality ratio, QR:

$$\mathrm{QR}(\mathcal{B}_j; \mathcal{S}_{\mathcal{I}_1}, \mathcal{S}_{\mathcal{I}_2}) = \frac{\mathcal{Q}(\mathcal{B}_j; \mathcal{S}_{\mathcal{I}_1})}{\mathcal{Q}(\mathcal{B}_j; \mathcal{S}_{\mathcal{I}_2})}. \tag{A.4}$$

In order to assess if two sequence variants, s_i and s_k, reflect the same underlying DNA sequence, we compute the probability that s_i is equivalent to s_k. We propose calculating the probability of a random sequence, $\mathcal{S}_k$, being a variation of the sequence s_i (using the equivalence notation "$\equiv$") by combining the coverage of the two sequences and the quality in and around the bases, where the two sequences mismatch. That is, the product of probability ratios, $P_k(j)$, over the mismatching bases, weighted by our belief, w_i, in the sequence, s_i, compared to s_k, given the following information:

(1) The two sequences are not equal.

(2) We know the set indices, where the two sequences mismatch, $\mathcal{I}_M$.

(3) The probability of error at and around the mismatching bases, $\mathcal{P}_{\mathcal{I}_M}$.

Thus, we propose

$$\mathbb{P}\left(\mathcal{S}_k \equiv s_i | s_k \neq s_i, \mathcal{I}_M, \mathcal{P}_{\mathcal{I}_M}\right) = w_i \prod_{j \in \mathcal{I}_M} P_k(j). \tag{A.5}$$

Let φ_i and φ_k be the coverage of sequence i and k, respectively. The weights, w_i, and the probability ratios, $P_k(j)$, were given as:

$$w_i = \frac{\varphi_i}{\varphi_k + \varphi_i},$$

with $\mathcal{P}_{kh}$ being the h'th error probability of the k'th sequence

$$P_k(j) = \frac{\max\limits_{h \in \partial(j)} \left\{ \dfrac{\mathcal{P}_{kh}}{|h-j|+1} \right\}}{\sum\limits_{h \in \partial(j)} \dfrac{\mathcal{P}_{kh}}{|h-j|+1}},$$

where $\partial(j)$ is the neighbourhood bases of j, i.e. defining j's neighbours at distance t:

$$\partial(j) = \{h | h \in [j-t; j+t]\},$$

where $\partial(j)$ was trimmed if j was close to the start or end of the sequence. In accordance with the Ion Torrent workflow, we set $t = 5$. That is, $\partial(j)$ is base j and its 10 closest neighbouring bases (5 to each side).

Eq. (A.5) should be interpreted as the probability of $\mathcal{S}_k$ being a variation of s_i. It follows that it creates an asymmetric matrix of probabilities

$$\mathbb{P}\left(\mathcal{S}_k \equiv s_i | s_k \neq s_i, \mathcal{I}_M, \mathcal{P}_{\mathcal{I}_M}\right) \neq \mathbb{P}\left(\mathcal{S}_i \equiv s_k | s_i \neq s_k, \mathcal{I}_M, \mathcal{P}_{\mathcal{I}_M}\right).$$

It should be noted that, as Eq. (A.5) includes weighted probabilities, the sum over all k's may differ from 1.

2.3.3 Heterozygote imbalance

We examined the variability of the coverage between the two alleles of a heterozygous marker by analysing the heterozygote balance, $\mathrm{H_b}$.

When shown against the average coverage, $\bar{\varphi}$ (see Eq. (A.8)), we compared it with the heuristic limits 0.6 and $1/0.6$ for CE [20]. We defined the heterozygote balance, H_b, as

$$H_b = \frac{\varphi_H}{\varphi_L}, \tag{A.6}$$

where φ_H and φ_L represent the coverage of the high and low molecular weight alleles, respectively. H_b contains more information than e.g. $H'_b = \min\{\varphi_H, \varphi_L\} / \max\{\varphi_H, \varphi_L\}$ and is therefore preferred [23].

We analysed the frequency and coverage of sequences identified as originating from the complementary strand to help explain imbalances in coverage between the alleles of a heterozygous marker. That is, we were not interested in large differences between the coverage of forward and complementary strands of a specific allele in itself, but the difference of the difference between the coverage of forward and complementary strands of the two alleles on a heterozygous marker.

The coverage pertaining to the forward and complementary strands were denoted as φ^F and φ^R, respectively. Let $H_b^F = \varphi_H^F / \varphi_L^F$, i.e. the heterozygote balance where only the coverage of the forward strand was used. The heterozygote balance was de-constructed as follows:

$$H_b = \frac{\varphi_H}{\varphi_L} = \frac{\varphi_H^F + \varphi_H^R}{\varphi_L^F + \varphi_L^R} = \frac{1 + \dfrac{\varphi_H^R}{\varphi_H^F}}{1 + \dfrac{\varphi_L^R}{\varphi_L^F}} \frac{\varphi_H^F}{\varphi_L^F} = \frac{1 + \varphi_H^R / \varphi_H^F}{1 + \varphi_L^R / \varphi_L^F} H_b^F,$$

dividing by H_b^F and taking the logarithm on both sides of the equality yields the following relation:

$$\log\left(\frac{H_b}{H_b^F}\right) = \log\left(\frac{1 + \varphi_H^R / \varphi_H^F}{1 + \varphi_L^R / \varphi_L^F}\right). \tag{A.7}$$

If the coverage of the forward and complementary sequences of the two alleles were perfectly balanced, it would be zero. The left-hand side of Eq. (A.7) is an approximate measure of the percentage difference of H_b when the coverage of the complementary strands were excluded. If the absolute value of the left- and right-hand side were large simultaneously, we attributed the change in H_b to the difference in coverage of the forward and complementary strands between the two alleles on a heterozygous marker.

2.3.4 Signal stability

In CE, the average peak height/area was used as an estimate of the amount of DNA [24]. We examined the behaviour of the average coverage as the amount of DNA was decreased. The average coverage of a marker, m, was given as

$$\bar{\varphi}_m = \frac{\sum_a \varphi_{ma}}{2}, \tag{A.8}$$

where φ_{ma} is the coverage of allele a of marker m. The sum in the numerator is taken over all alleles a on marker m. If the marker was unimportant or clear from the context, we dropped the subscript.

2.3.5 Stutters

We were interested in the rate of allele stutters. The stutter ratio (and proportion) has been shown to increase as the parental allele length increases if the repeat structure of the parental allele is simple [25, 26]. If the repeat structure is compound, complex, or has microvariants, stutters must be treated differently than stutters of simple repeat structures. It has been proposed that the longest uninterrupted stretch, LUS, is a better predictor of the stutter ratio than parental allele length [27].

This hypothesis was tested as the sequences are known in MPS and the LUS information is obtained. We defined the stutter ratio as

$$\mathrm{SR} = \frac{\varphi_{\mathrm{Stutter}}}{\varphi_{\mathrm{Parent}}}, \tag{A.9}$$

where $\varphi_{\mathrm{Stutter}}$ and $\varphi_{\mathrm{Parent}}$ are the coverages of the stutter and parent alleles, respectively.

Linear regression models using both allele repeat lengths and LUS as predictors were fitted and compared:

$$\mathrm{SR} = \beta_0 + \beta_1 X, \tag{A.10}$$

where β_0 and β_1 are marker dependent, and X is either the allele length or the LUS. We compared the linear model fits using R^2-values for numerical comparison, and we fitted a generalised additive model [28], a flexible non-linear, smooth model, for visual comparison.

2.3.6 Shoulders

We were interested in the frequency with which insertions and deletions occurred. In order to analyse this frequency, we examined the coverage of the shoulders surrounding an allele and defined the shoulder ratio analogous to Eq. (A.9), i.e.

$$\mathrm{SR}' = \frac{\varphi_{\mathrm{Shoulder}}}{\varphi_{\mathrm{Parent}}}, \tag{A.11}$$

where $\varphi_{\mathrm{Shoulder}}$ may refer to both left and right shoulders. In contrast to stutter, prediction of the shoulder ratio is not commonplace. Therefore, we examined the relationship between the parent allele lengths and the shoulder ratios.

2.3.7 Non-systematic errors

We observed a large number of sequences with low coverage. Although these sequences may have been the products of insertions, deletions, substitutions, or stutters (or even multiple combinations thereof), we found no association with known artefacts. Therefore, we categorised them as non-systematic errors or noise.

The simplest method to handle the remaining noise would be to remove the sequences as is commonly done with CE, where peaks lower than 50 RFU (relative fluorescence units) are removed from consideration by applying thresholds.

A naïve approach to create such thresholds is to base it on the percentage, p, of the total marker coverage or the coverage of the most prevalent sequence, $\mathcal{C}$, i.e.

$$\mathcal{T}_{\mathrm{naïve}} = p \cdot \mathcal{C}. \tag{A.12}$$

Applying static thresholds (which may be dependent marker) was not considered a viable option due to the imbalance in allele coverages among markers.

We propose fitting the distribution of the coverage with the aim of recognising more of the systematic errors for the analysis of samples with DNA mixtures as is seen in the probabilistic models applied to CE, see e.g. [29]. We fitted the distribution using both a one-inflated negative binomial model, OINB [30] (an extension of the zero-inflated

count models [31]) and a geometric model. The OINB model was defined, as follows:

$$\text{OINB}(x;\theta,\lambda,\pi) = \begin{cases} \pi + (1-\pi)f(1;\theta,\lambda) & \text{if } x = 1 \\ (1-\pi)f(x;\theta,\lambda) & \text{if } x > 1, \end{cases} \tag{A.13}$$

where x is a positive integer (here the coverage), λ is the mean value parameter, θ is the shape parameter, π is the mixture parameter, and $f(\cdot;\theta,\lambda)$ is the zero-truncated negative binomial distribution. The one-inflation was added to the negative binomial model as preliminary analyses showed a larger number of sequences with a coverage of one, when compared to the number of sequences with a coverage of two.

The threshold was then based on a quantile, q, and the theoretical standard deviation, σ, of the fitted distribution:

$$\mathcal{T} = q + 3\sigma, \tag{A.14}$$

where q is based on the proportion p, which was chosen such that $\mathcal{T}$ isolated the extreme values of the distribution, i.e. the alleles and stutters. Note, that the scalar 3 was chosen arbitrary.

2.3.8 Simulating data samples

When analysing the signal variability, more data is needed. Therefore, we simulated sequence coverage data in accordance with peak height simulations made in CE. The simulations were performed using an extension of the classic binomial sampling scheme [32]. The simulated data was specifically needed in order to determine possible predictors of the amount of template DNA. We extended the binomial sampling scheme as follows: Given n copies of template DNA and the size of the chip N, the sampling process was broken down as follows:

(1) The amount of DNA extracted for PCR amplification:

$$n_0 = \text{Bin}\left(n, \pi_{\text{aliquot}}\right).$$

(2) The number of DNA copies in cycle t:

$$n_t = n_{t-1} + \text{Bin}\left(n_{t-1}, \pi_{\text{PCReff}}\right).$$

(3) Given $n_{\text{IBA}} = N/(\#\text{ of IBA's})$, the total number of slots per IBA, the amount of available slots was defined as:

$$\text{IBA}_{\text{chip}} = \Gamma\left(n_{\text{IBA}}, 1\right).$$

The scale parameter was set equal to 1 as it implies:

$$\mathbb{E}\left[\text{IBA}_{\text{chip}}\right] = n_{\text{IBA}}.$$

(4) If a IBA contained multiple markers, we defined the number of available slots per marker using a multinomial distribution as:

$$\textbf{Marker}_{\text{IBA}} = \text{Mult}(\text{IBA}_{\text{chip}}, \boldsymbol{\pi}_{\text{MarkerEff}}),$$

assigning the slots to the alleles on a heterozygous marker.

3 Results

3.1 Inclusion of reverse complementary sequences

In order to identify the STR regions, as described in Section 2.3.1, we created plots similar to those of CE, as seen in Fig. A.1 panel (A), showing the coverage against the length of the repeat region of marker D3S1358 of a 2 ng sample. The figure shows that the individual had genotype 17, 18 at D3S1358, which is the correct genotype. Fig. A.1 panel (B) shows the same marker of the same sample. However, only reads identified in the forward read direction were included. The figure points to the genotype of 17, 17. The reduction in coverage of alleles 17 and 18 were 4, 5% and 94%, respectively (i.e. the fraction of reads originating from the reverse complementary strands were 4.5% and 94% for alleles 17 and 18, respectively).

In order to identify similar heterozygote imbalances caused by imbalances in the coverages of the fractions of forward and reverse complementary strands between the two alleles of a heterozygote, we plotted the left- and right-hand side of Eq. (A.7) against each other for all autosomal markers in the 10plex, Fig. A.2. We observed large imbalances at the coverages of the alleles of D3S1358 and D5S818. D3S1358 showed extreme values for one of the dilution series for contributor 2. This imbalance was further illustrated in Fig. A.3 showing the fraction of complementary strands for one of the extreme values.

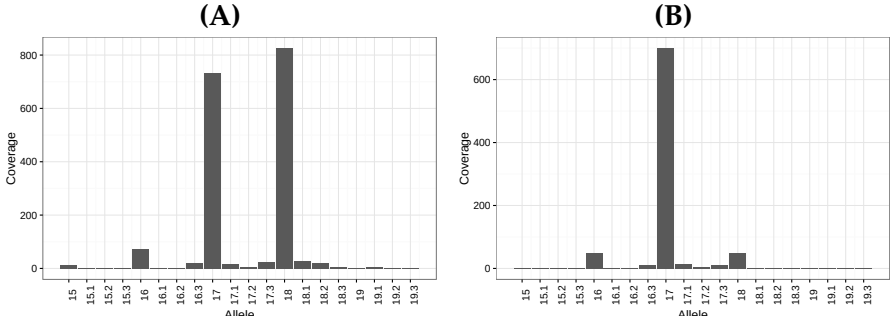

Fig. A.1: Panel (A) and (B) shows all unique sequences of marker D3S1358 of a 2ng sample, aggregated by repeat length. Furthermore, panel (B) has been restricted to sequences identified in the forward read direction.

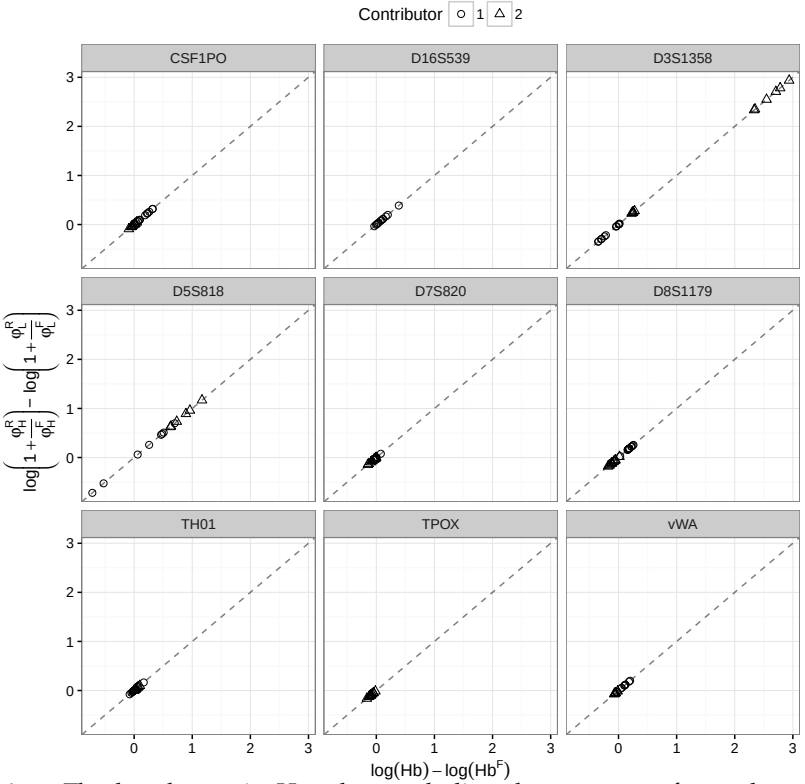

Fig. A.2: The log-change in H_b when excluding the coverage of complementary strands against the log-difference in coverage pertaining to the forward and reverse complementary strands between the alleles of a heterozygote. The two contributors of the samples were indicated using ○ and △, respectively.

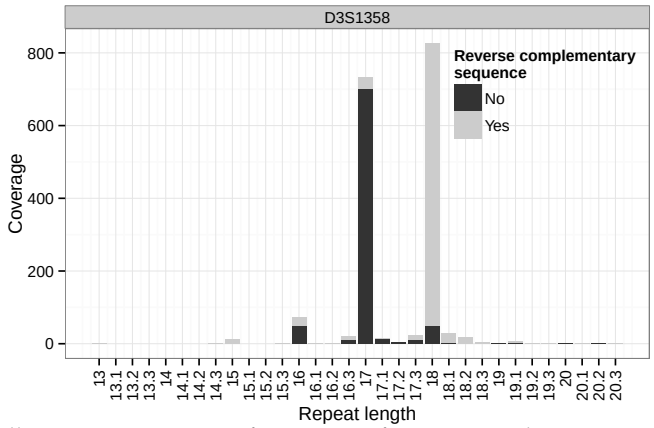

Fig. A.3: All unique sequences of D3S1358 of a 2ng sample, aggregated by repeat length and split into sequences identified as belonging to the reverse complementary strand or not shown in gray and black, respectively.

We, therefore, unless otherwise stated, summed the coverage of the forward and complementary strands to avoid inducing this kind of imbalance.

3.2 The quality

In analysing the sequence quality, we did not include the complementary sequences, as they were sequenced in opposite direction and they could, therefore, not had been analysed simultaneously with the non-complementary sequences. In order to understand why, assume we had sequenced a read and its complementary and recall from Section 2.3.2 that the quality is decreasing with each sequenced base. If we wanted to aggregate the information, as when calculating the base quality in Eq. (A.3), then we would have reversed one of the reads and taking its complementary and, therefore, its quality sequence would have been reversed. Assuming we reversed the complementary read, the quality would no longer be decreasing, but increasing; the consequence being that when aggregating the base qualities, the quality would become constant, as one read had decreasing quality, while the other had increasing quality.

We excluded the reverse complementary strands when the quality was analysed. We could have analysed them separately, but the resulting analysis would have been equivalent to that of the forward

strands.

3.2.1 The qualities of sequences

The sequence quality of the 40 most prevalent sequences of repeat length 11 and 12 of the D16S539 marker of a 2 ng sample is seen in Fig. A.4. In both cases, the most prevalent sequence belonged to the contributor of the sample and the remaining sequences were classified as basecall errors of the most prevalent sequence. The figure shows a uniform median quality across the 40 most prevalent sequences for both alleles. We have only shown the results of a 2 ng sample, as we saw similar results for the remainder of samples in the dilution series.

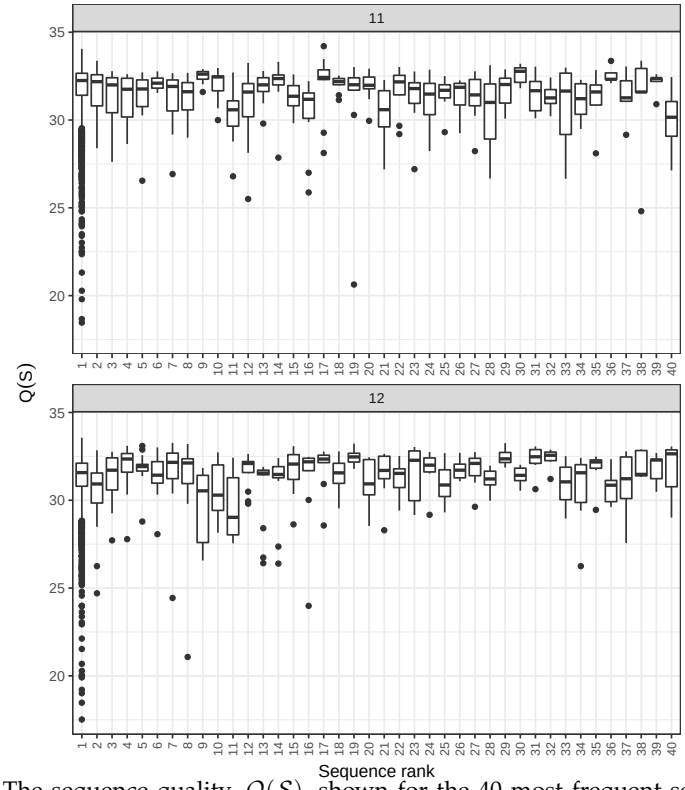

Fig. A.4: The sequence quality, $\mathcal{Q}(\mathcal{S})$, shown for the 40 most frequent sequences of repeat length 11 and 12 of the D16S539 marker of a 2ng sample. The 40 most frequent sequences were shown in order of prevalence, called their sting rank.

3.2.2 The quality ratios of bases

Fig. A.5 shows the base quality ratio for each of the 2-10 most frequent sequences using the most frequent sequence as reference, i.e. D16S539 allele 12 of a 2 ng sample. The quality ratio dropped at or around the base, where a substitution had occurred (indicated by the bullet in the figure).

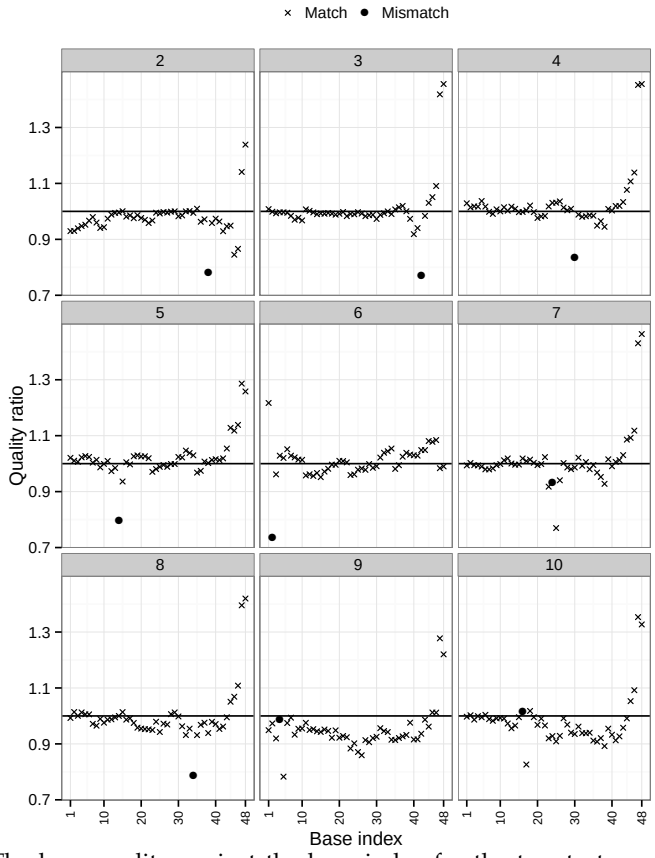

Fig. A.5: The base quality against the base index for the two to ten most frequent (as indicated above each subplot) of repeat length 12 of the D16S539 marker of a 2ng sample using the most frequent sequence as the reference.

3.2.3 Probability one sequence is a variation of another (based on base quality)

As an example, we considered the five most frequent sequences with a repeat length of 20 at the vWA STR locus. Table A.1 shows sequences in order of frequency, their coverage, and the number of base call errors when compared to the sequence of the most frequent sequence.

Table A.1: The probability matrix corresponding to the five most frequent sequences of vWA allele 20 of a 2 ng sample. The sequences $s_2, ..., s_5$ each had one mismatch when compared to s_1.

		$\mathcal{S}_k$					
		s_1	s_2	s_3	s_4	s_5	Coverage
s_i	s_1	–	0.3625	0.3018	0.3187	0.3683	11,626
	s_2	0.0022	–	0.0237	0.0303	0.0391	102
	s_3	0.0016	0.0169	–	0.0257	0.0406	77
	s_4	0.0014	0.0142	0.0157	–	0.0236	56
	s_5	0.0011	0.0105	0.0124	0.0172	–	52

From the frequencies in Table A.1, s_1 is far the most likely true sequence of the allele (most likely the allele of the victim in the case of a DNA mixture sample). Now, the question comes up: Are the remaining sequences, $s_2, \ldots, s_5$, simply variations caused by errors of the most frequent sequence? Using Eq. (A.5), we calculated that the probability of the sequence s_2 being variation of s_1 was 36.3%. We could also have asked the converse, i.e. if sequence s_1 is a variation of s_2, in which case the probability was 0.2%. The remaining comparisons can be found in Table A.1, where the value seen in row i of column k corresponds to the probability of sequence $\mathcal{S}_k$ being a variation of sequence s_i.

3.2.4 Preferential detection

The quality decreases, or conversely the probability of error increases, with each sequenced base, which had two consequences due to the method for identification of the STR regions:

(1) The regions were typically found in the beginning of the read, as

was illustrated in Figure A.6 showing the average beginning and end of D16S539 marker identified reads of a 2ng sample with repeat length 12.

(2) Longer alleles could be more difficult to detect. Thus, shorter alleles were preferentially detected, which could create a coverage imbalance between heterozygotes far apart from each other in length. The preferential detection was partly combated by allowing for mutations in the flanking regions; allowing for more mutations in the flanking region would make the imbalance smaller, but more erroneous reads would be retained.

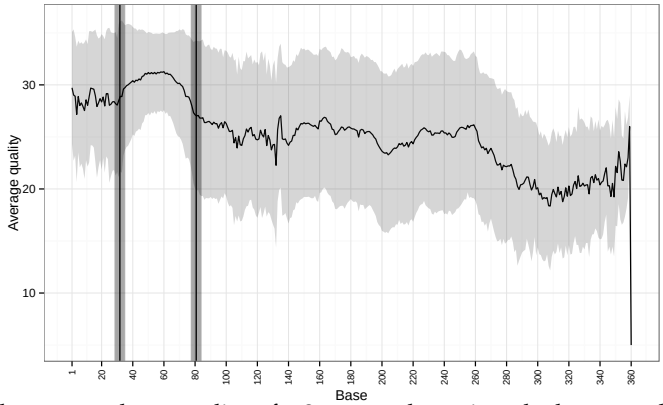

Fig. A.6: The average base quality of a 2ng sample against the base number. The two vertical bars show the average beginning and end position of all D16S539 allele 12. The shaded areas indicate the mean quality plus/minus two standard deviations.

3.3 Heterozygote imbalance

Fig. A.7 shows H_b against the average allele coverage. The dashed lines were set at 0.6 and 1/0.6 ([32]). The figure resembles similar plots from CE, an indication of the common underlying dynamics of the PCR amplification.

3.4 Signal stability

Plotting $\bar{\varphi}$ against the amount of input DNA (Fig. A.8) showed that $\bar{\varphi}$ is not an apt predictor of the amount of template DNA in contrast to

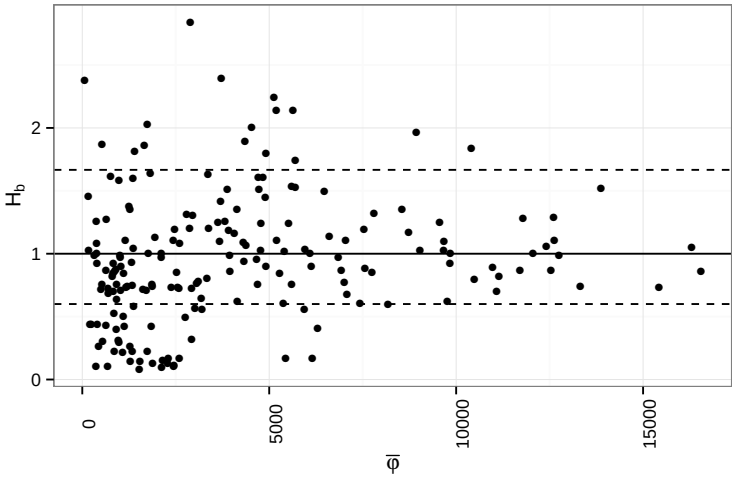

Fig. A.7: The heterozygote balance, H_b, plotted against the average allele coverage, $\bar{\varphi}$. The horizontal dashed lines indicate the heuristic limits 0.6 and 1/0.6, that are frequently used in CE.

the situation with CE [24]. This is a consequence of IonExpress Barcode adapter (IBA) normalisation. As MPS allows for multiple samples to be sequenced in parallel, the number of samples for each IBA were approximately equimolecular, thereby, normalising the number of strands for each IBA.

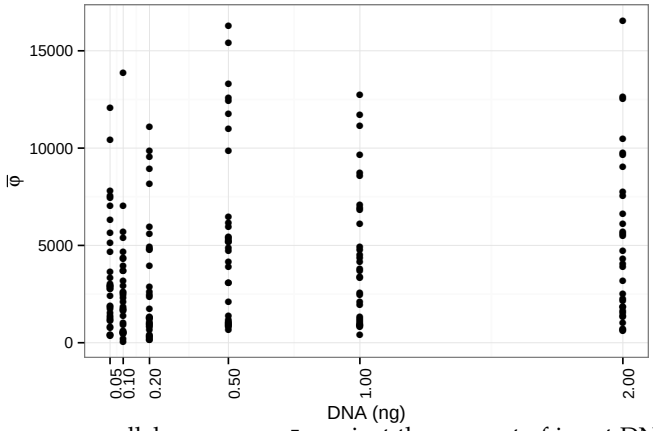

Fig. A.8: The average allele coverage, $\bar{\varphi}$, against the amount of input DNA measured in nanograms.

By applying the simulation scheme introduced in Section 2.3.8, we

simulated 1,000 dilution series of six dilutions, halving the amount of DNA each time with an initial amount of template DNA set at $n = 1,000$. Fig. A.9 and A.10 show that the simulated heterozygote balance had a similar shape as the real data.

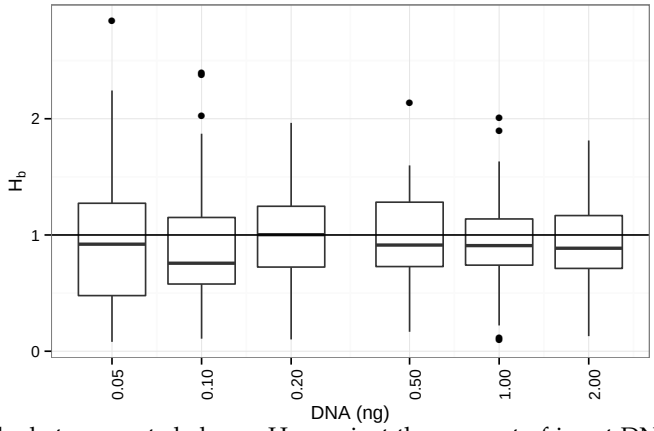

Fig. A.9: The heterozygote balance, H_b, against the amount of input DNA measured in nanogram for the real data. The abscissa is shown on a $\log_{10}$-scale.

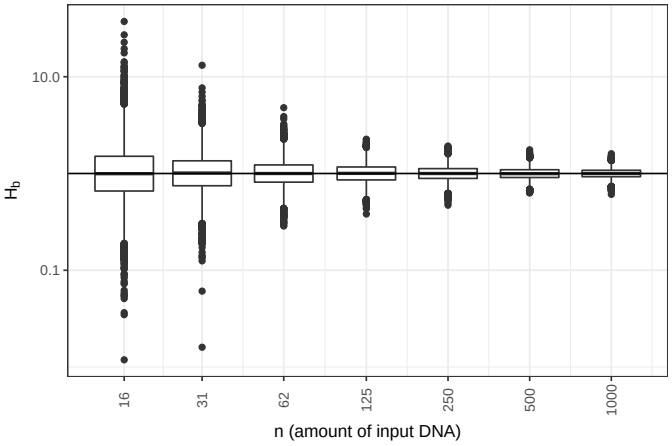

Fig. A.10: The heterozygote balance, H_b, against the amount of input DNA, for the simulated data. Both the abscissa and ordinate are shown on a $\log_{10}$-scale.

3.5 Stutters

Fig. A.11 shows a clear example of the difference between the two predictors of stutters at TH01. The R^2 value for the linear model fitted using allele lengths and LUS were 0.02 and 0.78, respectively. The LUS hypothesis was further supported by the non-linear fitted line, as it was laying directly on top of the linear one. The TH01 marker was chosen as it very clearly showed the differences between the two methods due to the microvariant at 9.3. In Fig. A.12 the stutter ratio is plotted against the parent allele length and LUS for the vWA marker. The R^2 values for the vWA marker were 0.58 and 0.64, respectively, using parental allele length and LUS as predictors. The remainder of the markers can be seen in Fig. A.13 and A.14 for allele lengths and LUS, respectively.

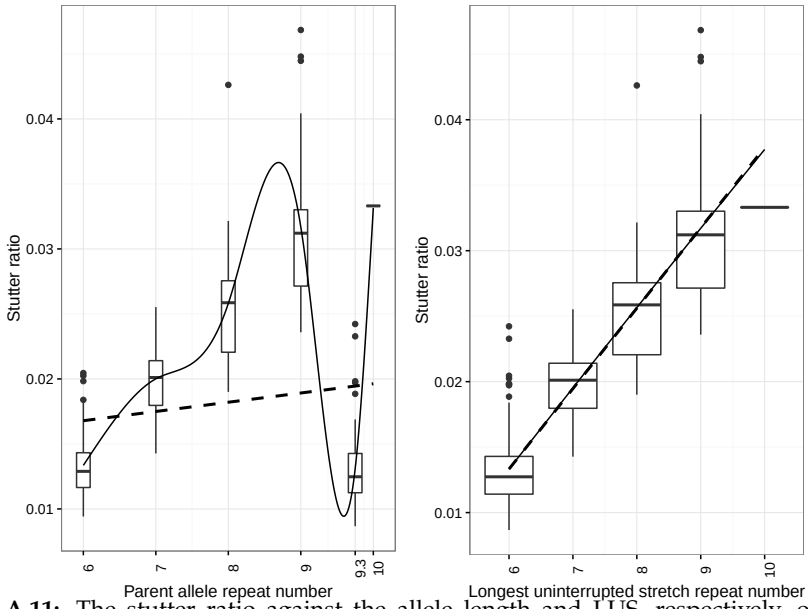

Fig. A.11: The stutter ratio against the allele length and LUS, respectively, of the TH01 STR marker. The dashed and solid lines represent linear model and generalised additive model fits, respectively. The boxplots were constructed such that the boxes showed the first, second, and third quartiles (Q1, median, and Q3, respectively), while the whiskers were defined as Q3 + 1.5×IQR and Q1 - 1.5×IQR for the upper and lower whisker, respectively (IQR $=$ Q3 $-$ Q1). Any point outside these whiskers were classified as outliers and indicated by dots. All boxplots followed the same structure [33].

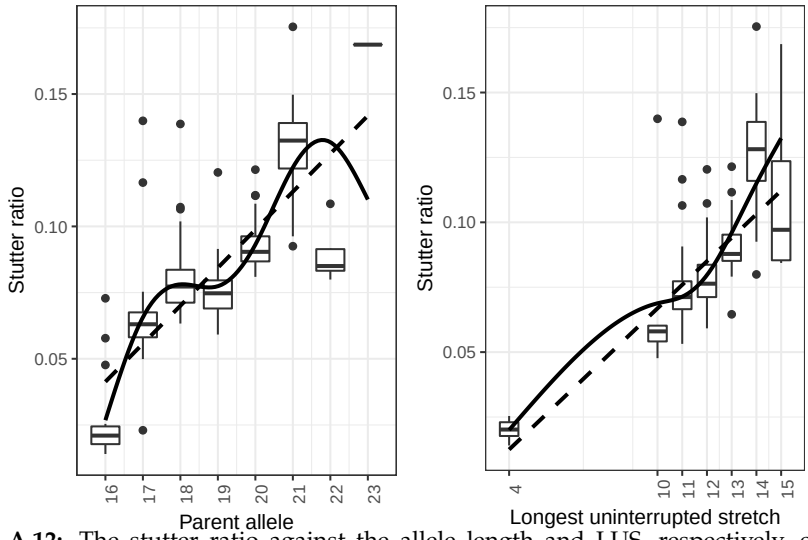

Fig. A.12: The stutter ratio against the allele length and LUS, respectively, of the vWA STR marker. The dashed and solid lines represent linear model and generalised additive model fits, respectively.

When comparing the remaining markers of the two figures, we saw a clear difference in the non-linear fitted (solid) lines for markers with a compound, complex, and micro-variant repeat structure. The LUS had a clear advantage as a linear-predictor of stutter-ratio compared to that of allele length, with the markers D3S1358 and TH01 showing the biggest improvement. However, as the markers in the HID STR 10-plex are fairly short and simple, the gain is not as large as it may be expected for more heterogeneous STRs.

3.6 Shoulders

Shoulders can be observed as sequences, one base shorter or longer, depending on whether a deletion or an insertion has occurred, respectively. The shoulders of e.g. allele 18, seen in Fig. A.1 panel (A) had the sequences of lengths 17.3 and 18.1, referred to as the allele's left and right shoulder, respectively.

As in the case of stutters, we were interested in the rate at which a shoulder was observed and it the length of the allele had any impact on this rate. We measured this rate by the shoulder ratio defined in Eq. (A.11). Plotting the shoulder ratio against the allele length (Fig. A.15)

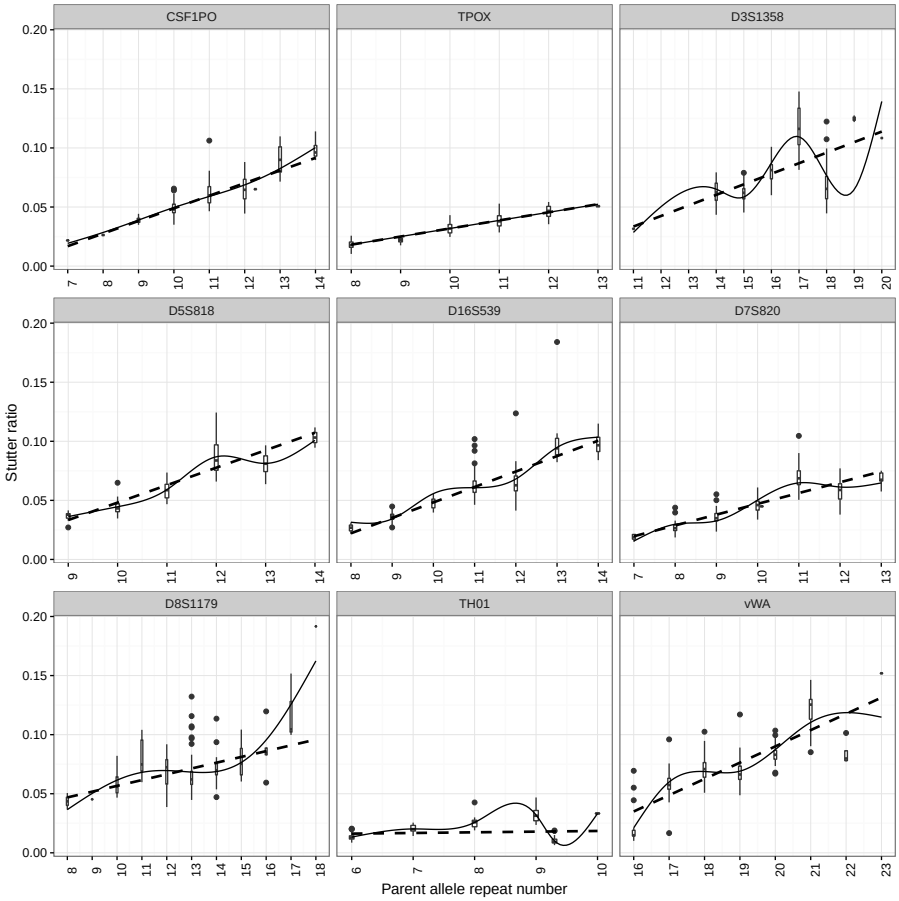

Fig. A.13: The stutter ratio against the allele length for all markers in the HID STR 10 plex. The dashed and solid lines represent linear model and generalised additive model fits, respectively.

showed that the shoulder ratio varied among the markers, though it was fairly stable for all allele lengths.

3.7 Non-systematic errors

The total number of drop-outs, drop-ins (total as well as sequence variations, stutters, and shoulders), adjusted drop-ins, marker drop-outs for the OINB, geometric, and naïve methods can be seen in Table A.2. The OINB method had two drop-outs and no adjusted drop-ins, while the geometric method had both a drop-out and an adjusted drop-in.

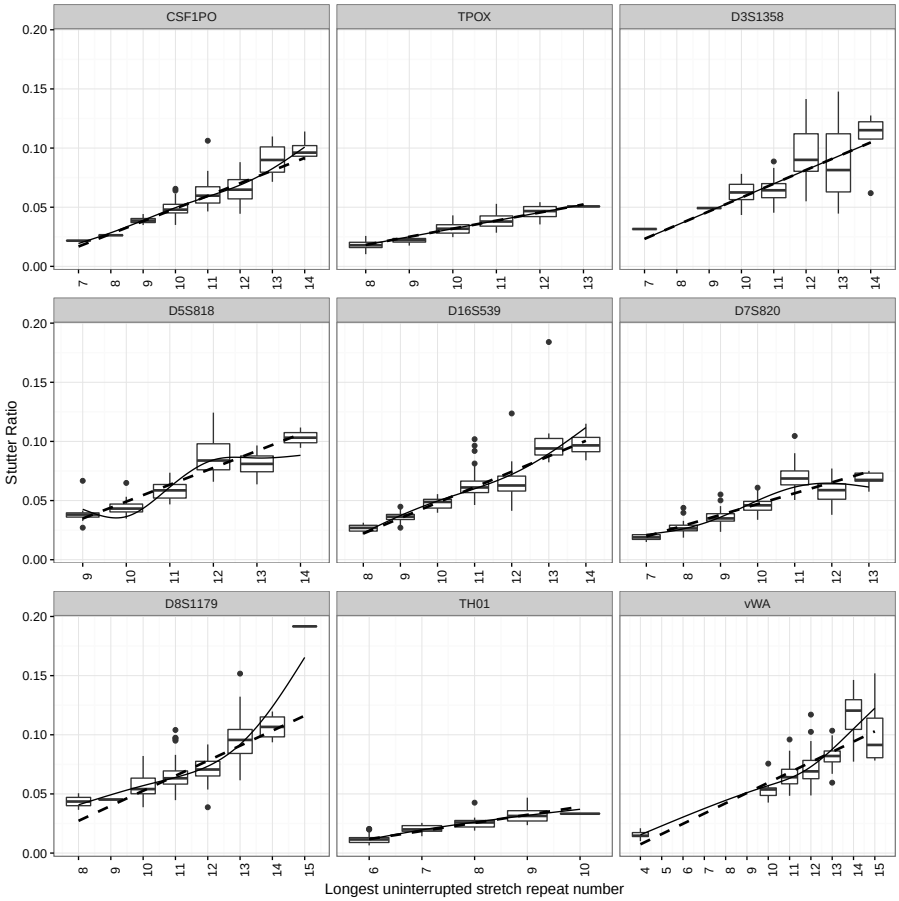

Fig. A.14: The stutter ratio against the LUS for all autosomal markers in the HID STR 10 plex. The dashed and solid lines represent linear model and generalised additive model fits, respectively.

The naïve threshold at 0.05 had no drop-out and adjusted drop-in, whereas the threshold of 0.1 had 12 drop-outs and no adjusted drop-in.

The OINB and geometric method retained on average 1.83 and 1.20 systematic errors per marker, respectively, for potential use in DNA mixture samples. The naïve thresholds at 0.05 and 0.1 on the other hand retained only 0.14 and 0.0048 systematic errors per marker, respectively. It follows that if the systematic errors are shown to be useful in MPS, as they are in CE, we would prefer the OINB or geometric method.

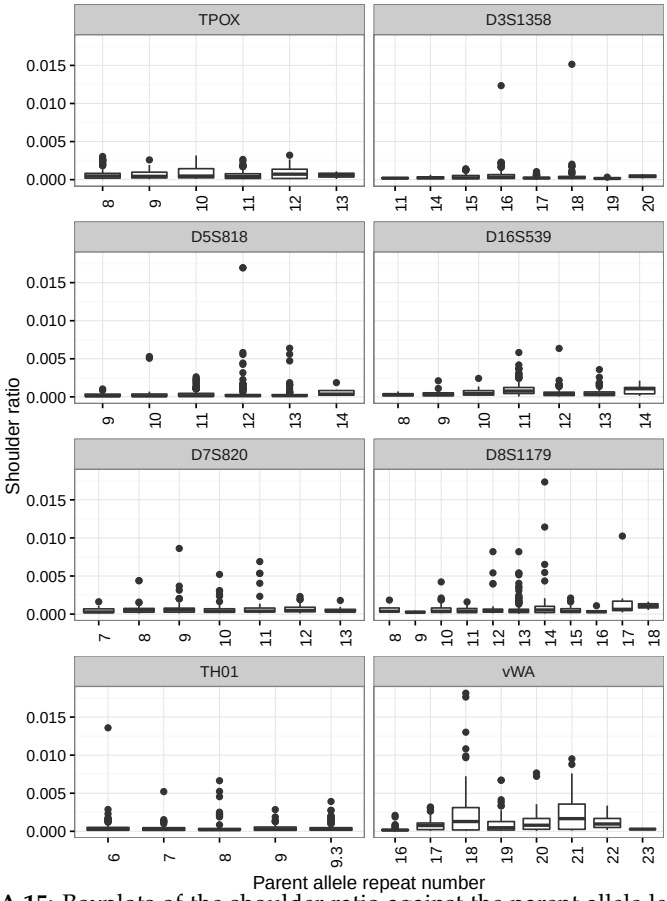

Fig. A.15: Boxplots of the shoulder ratio against the parent allele length.

Table A.2: The number of drop-outs, drop-ins (total, sequence variants, stutters, and shoulders), adjusted drop-ins, and total marker drop-outs for thresholds created using the OINB, geometric distributions, and naïve method (at 5% and 10%).

	Drop-out	Drop-in	Sequence variant Drop-in	Stutter Drop-in	Shoulder Drop-in	Adjusted Drop-in	Marker Drop-out
OINB	2	379	42	196	141	0	0
Geometric	1	250	32	156	61	1	0
Naïve 5%	0	30	0	9	21	0	0
Naïve 10%	12	1	0	0	1	0	0

4 Discussion

The extreme strand bias presented in Section 3.1 and illustrated in Fig. A.1 and A.3 was mainly included as a warning for others, as we could not specifically pinpoint the cause of this bias from the data at hand, and we have not observed a similar extreme in other data. Our best guess is that during the PCR process, secondary structures were created in one sequence, which were not created in the other sequence, and that these secondary structures were difficult for the polymerase to pass during sequencing, see e.g. [34, 35] for a more thorough discussion on this matter.

The usefulness of the sequence quality, $\mathcal{Q}(\mathcal{S})$, is debatable because of the shear number of bases in a sequence and the fact that our method of STR identification restricts the number of sequences based on the quality, which is why the more complicated asymmetric probability is introduced.

It is clear from the analysis of stutter ratio that the LUS is a better linear predictor than the allele repeat length, see Fig. A.13 and A.14. However, the markers from the HID STR 10-plex panel are, compared to markers used in CE, short and the repeat patterns are not as complicated. Thus, the hypothesis should be investigated using compound, complex, and micro-variant STR markers.

The plot of the heterozygote balance against the average allele coverage (Fig. A.7) resembles its counterpart in CE. However, to get a clearer picture, more data is needed.

A simple consequence of IBA normalisation (and subsequent normalisation) implied that we were unable to estimate the amount of template DNA using the average coverage directly. Calculating the probability of drop-out, therefore, becomes more complicated.

By examining Fig. A.9, A.10, and A.16, we saw that the variation of the heterozygote balance increased as the amount of template DNA was decreased. We, therefore, hypothesise that the standard deviation of the heterozygote balance could be used as an estimate of the amount of template DNA. The increase in standard deviation is a direct consequence of binomial sampling and, therefore, the hypothesis should hold for both MPS and CE. Furthermore, if the distribution of the coverage was assumed to follow a gamma distribution, as the peak height in [29], the hypothesis would follow directly from the choice of

distribution. This hypothesis was not tested here, but is offered for future research.

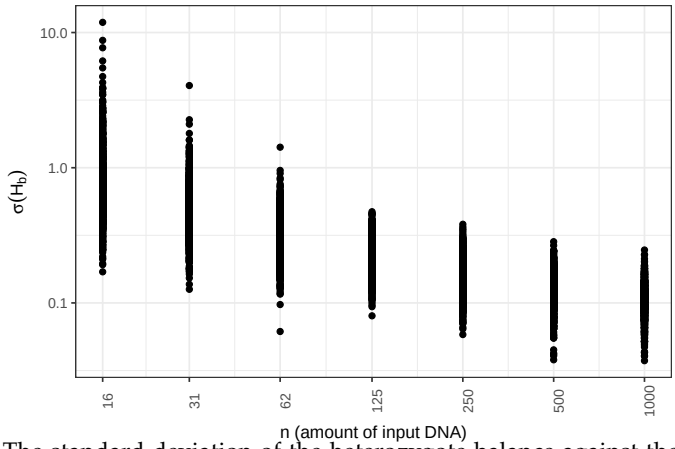

Fig. A.16: The standard deviation of the heterozygote balance against the amount of input DNA, for the simulated data. Both the abscissa and ordinate are shown on $\log_{10}$-scales.

When applying the noise threshold, we aimed to preserve some of the systematic errors as we assume that they will be useful as they are in CE when mixture samples are going to be investigated with MPS in the future.

References

[1] K. Lazaruk, P. S. Walsh, F. Oaks, D. Gilbert, B. B. Rosenblum, S. Menchen, D. Scheibler, H. M. Wenz, C. Holt, and J. Wallin, "Genotyping of forensic short tandem repeat (str) systems based on sizing precision in a capillary electrophoresis instrument," *Electrophoresis*, vol. 19, no. 1, pp. 86–93, 1998. [Online]. Available: http://dx.doi.org/10.1002/elps.1150190116

[2] J. Butler, *Fundamentals of Forensic DNA Typing*. Academic Press, 2009.

[3] S. Fordyce, M. Avila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F. Petersen, E. Willerslev, A. Hansen, N. Morling, and M. Gilbert, "High-throughput sequencing of core STR loci

for forensic genetic investigations using the Roche Genome Sequencer FLX platform," *BioTechniques*, vol. 51, pp. 127 – 133, 2011.

[4] D. M. Bornman *et al.*, "Short-read, high-throughput sequencing technology for STR genotyping." *BioTechniques*, pp. 1–6, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22668513

[5] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, and D. Deforce, "Forensic STR analysis using massive parallel sequencing," *Forensic Science International: Genetics*, vol. 6, no. 6, pp. 810–818, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2012.03.004

[6] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich, "lobSTR: A short tandem repeat profiler for personal genomes," *Genome Research*, vol. 22, no. 6, pp. 1154–1162, 2012.

[7] D. H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, J. L. King, and B. Budowle, "STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data," *Forensic Science International: Genetics*, vol. 7, no. 4, pp. 409–417, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2013.04.005

[8] S. Y. Anvar, K. J. van der Gaag, J. W. F. van der Heijden, M. H. A. M. Veltrop, R. H. A. M. Vossen, R. H. de Leeuw, C. Breukel, H. P. J. Buermans, J. S. Verbeek, P. de Knijff, J. T. den Dunnen, and J. F. J. Laros, "Tssv: a tool for characterization of complex allelic variants in pure and mixed genomes," *Bioinformatics*, vol. 30, no. 12, p. 1651, 2014. [Online]. Available: +http://dx.doi.org/10.1093/bioinformatics/btu068

[9] E. Rockenbauer, S. Hansen, M. Mikkelsen, C. Børsting, and N. Morling, "Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing," *Forensic Science International: Genetics*, vol. 8, no. 1, pp. 68–72, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2013.06.011

[10] S. Dalsgaard, E. Rockenbauer, A. Buchard, H. S. Mogensen, R. Frank-Hansen, C. Børsting, and N. Morling, "Non-uniform phenotyping of D12S391 resolved by second generation sequencing," *Forensic Science International: Genetics*, vol. 8, no. 1, pp. 195–199, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2013.09.008

[11] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Børsting, and N. Morling, "Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles," *Forensic Science International: Genetics*, vol. 12, pp. 38–41, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2014.04.016

[12] M. Scheible, O. Loreille, R. Just, and J. Irwin, "Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers," *Forensic Science International: Genetics*, vol. 12, pp. 107–119, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2014.04.010

[13] C. Van Neste, M. Vandewoestyne, W. Van Criekinge, D. Deforce, and F. Van Nieuwerburgh, "My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing," *Forensic Science International: Genetics*, vol. 9, no. 1, pp. 1–8, 2014.

[14] D. H. Warshauer, J. L. King, and B. Budowle, "STRait Razor v2.0: The improved STR Allele Identification Tool-Razor," *Forensic Science International: Genetics*, vol. 14, pp. 182–186, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2014.10.011

[15] C. Børsting and N. Morling, "Next generation sequencing and its applications in forensic genetics," *Forensic Science International: Genetics*, vol. 18, pp. 78 – 89, 2015.

[16] S. Fordyce, H. Mogensen, C. B. rsting, R. Lagacé, C.-W. Chang, N. Rajagopalan, and N. Morling, "Second-generation sequecing of forensic STRs using the Ion Torrent^{TM} HID STR 10-plex and the Ion PGM," *Forensic Science International: Genetics*, vol. 14, pp. 132 – 140, 2015.

[17] S. L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, and N. Morling, "Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs," *Forensic Science International: Genetics*, vol. 21, pp. 68–75, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2015.12.006

[18] "FASTQ-format," https://en.wikipedia.org/wiki/FASTQ_format, accessed: 2016-05-09.

[19] "Torrent Suite technical notes and whitepaper," http://129.130.90.13/ion-docs/Technical-Notes-and-Whitepapers_6128100.html, accessed: 2016-12-14.

[20] P. Gill, R. Sparkes, and C. Kimpton, "Development of guidelines to designate alleles using an STR multiplex system," *Forensic Science International*, vol. 89, pp. 185 – 197, 1997.

[21] "Ion torrent technical notes - the per-base quality score system," http://129.130.90.13/ion-docs/Technical-Note---Quality-Score_6128102.html, accessed: 2016-12-14.

[22] B. Ewing and P. Green, "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities," *Genome Research*, vol. 8, pp. 186 – 194, 1998.

[23] H. Kelly, J.-A. Bright, J. Curran, and J. Buckleton, "The interpretation of low level DNA mixtures," *Forensic Science International: Genetics*, vol. 6, pp. 191 – 197, 2012.

[24] T. Tvedebrink, H. Mogensen, M. Stene, and N. Morling, "Performence of two 17 locus forensic identification STR kits – Applied Biosystem's AmpFlSTR® NGMSElect™ and Promega's PowerPlex® ESI17 kits," *Forensic Science International: Genetics*, vol. 6, pp. 523 – 531, 2012.

[25] M. Klintschar and P. Wiegand, "Polymerase slippage in relation to the uniformity of tetrameric repeat stretches," *Forensic Science International*, vol. 135, pp. 163 – 166, 2003.

[26] P. Walsh, N. Fildes, and R. Reynolds, "Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA," *Nucleic Acids Research*, vol. 24, pp. 2807 – 2812, 1996.

[27] C. Brookes, J. Bright, S. Harbison, and J. Buckleton, "Characterising stutter in forensic STR multiplexes," *Forensic Science International: Genetics*, vol. 6, pp. 58 – 63, 2012.

[28] T. Hastie and R. Tibshirani, "Generalized Additive Models," *Statistical Science*, vol. 1, pp. 297 – 318, 1986.

[29] R. Cowell, T. Graversen, S. Lauritzen, and J. Mortera, "Analysis of Forensic DNA Mixtures with Artefacts," *Royal Statistical Society. Journal Series C: Applied Statistics*, vol. 64, pp. 1 – 32, 2015.

[30] S. Vilsen, T. Tvedebrink, H. Mogensen, and N. Morling, "Modelling noise in second generation sequencing forensic genetics STR data using a one-inflated (zero-truncated) negative binomial model," *Forensic Science International: Genetics Supplement Series*, 2015.

[31] H. Lim, W. Li, and P. Yu, "Zero-inflated Poisson regression mixture model," *Computational Statistics and Data Analysis*, vol. 71, pp. 151 – 158, 2014.

[32] P. Gill, J. Curran, and K. Elliot, "A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci," *Nucleic Acids Research*, vol. 33, pp. 632 – 643, 2005.

[33] J. W. Tukey, *Exploratory Data Analysis*, Tukey, J. W., Ed. Addison-Wesley, 1977.

[34] F. Meacham, D. Boffelli, J. Dhahbi, D. Martin, M. singer, and L. Pachter, "Identification and correction of systematic error in high-troughtput sequencing data," *BMC Bioinformatics*, vol. 12, 2011.

[35] R. Ekblom, L. Smeds, and H. Ellegren, "Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria," *BMC Genomics*, vol. 15, 2014.

Paper B

Stutter analysis of complex STR MPS data

Søren B. Vilsen, Torben Tvedebrink, Poul Svante Eriksen, Claus Børsting, Christian Hussing, Helle Smidt Mogensen, and Niels Morling

The paper has been published in the
Forensic Science International: Genetics, vol. 35, pp. 107–112, 2018.

The layout has been revised.

Abstract

Stutters are common and well documented artefacts of amplification of short tandem repeat (STR) regions when using polymerase chain reaction (PCR) occurring as strands one or more motifs shorter or longer than the parental allele. Understanding the mechanism and rate by which stutters are created is especially important when the samples contain small amounts of DNA or DNA from multiple contributors. It has been shown that there is a linear relationship between the longest uninterrupted stretch (LUS) and the stutter ratio. This holds if there is only a single type of stutter variant. However, with massively parallel sequencing (MPS), we see that alleles may create different stutters corresponding to stuttering of different parts of the parental allele. This calls for a refinement of the LUS concept.

We analysed all uninterrupted stretches, here called blocks, and identified the block from which the stutter originated. We defined the block length of the missing motif (BLMM) as the length of the identified block. We found that the relationship between the stutter ratio and BLMM was linear using a simple system of recurrence relations. We found that the mean square error decreased by a factor upto 17.5 for compound and complex autosomal markers when using BLMM instead of LUS.

1 Introduction

DNA at a crime scene is often found in very small quantities. In forensic genetics, short tandem repeat (STR) regions are amplified using a polymerase chain reaction (PCR) [1]. A common artefact of amplifying STR regions is stuttering [2–7]. When a DNA strand is copied, the polymerase enzyme can skip (or repeat) a motif, denoted stutter (and back-stutter). As stuttering is a common artefact, it follows that predicting the rate of stuttering is important for interpretation of DNA profiles. Because massively parallel sequencing (MPS) also depends on the PCR process, it follows that understanding and predicting the rate of stuttering is important for the modelling and interpretation of MPS STR DNA mixture samples.

We measured the rate of stuttering by the stutter ratio (or stutter proportions), defined as:

$$S_R = \frac{y_a}{y_A} \qquad \left(\text{or } S_P = \frac{y_a}{y_A + y_a}\right), \tag{B.1}$$

where y_a and y_A are the coverages (the MPS analogue of the peak height/width in capillary electrophoresis (CE)) of the stutter and parental allele, respectively.

The hypothesis is that the more repetitive a strand is, the more likely the PCR process is to stutter (and, thus, to increase the stutter ratio). The simplest measure of the repetitiveness of a strand is its length. It has been shown that the relationship between the stutter ratio and the allele length is linear when the structure of the STR region is simple [4, 7]. However, this is not necessarily true if the structure is compound or complex. This led to the introduction of the LUS, defined as the longest uninterrupted stretch of repeated motifs within the allele [4]. Thus, the LUS works under the same hypothesis as the allele length. However, it takes into account that the probability of stuttering 'resets' when changing from one motif to another in compound or complex strands. It has been shown, that the relationship between the stutter ratio and LUS is linear in (1) capillary electrophoresis [4, 6–9] and (2) massively parallel sequencing (MPS), if only the most prominent stutter was considered [10].

With MPS, we can differentiate between different stutters of the

same parental allele. Consider the allele

$$[\texttt{AATG}]_{10}[\texttt{ACTT}]_4, \tag{B.2}$$

and assume that the following two possible stutters of the allele were observed:

$$[\texttt{AATG}]_9[\texttt{ACTT}]_4 \text{ and } [\texttt{AATG}]_{10}[\texttt{ACTT}]_3.$$

Note that all strands written in this paper follows the notation recommended by the ISFG DNA commission [11].

As they are both possible stutters of Eq. (B.2), they have the same LUS (i.e. 10). However, it is hypothesised that the stutter ratio of the second stutter would be much smaller than that of the first stutter as the second stutter has lost a motif from a much shorter stretch of the allele.

We propose a new predictor of stuttering, which we call the block length of the missing motif (BLMM). The BLMM determines the length of the stretch from where a motif has been lost. Thus, the BLMM works according to the same principles as the LUS but utilises the extra information obtained from sequencing of the DNA sample. We have restricted the analysis in this paper to consider only single stutters, but note that the BLMM concept could be extended to any type of stutter.

2 Materials and methods

2.1 Data

DNA was extracted from blood samples and buccal swabs collected on FTA cards from 366 individuals. DNA libraries were built using the ForenSeq™ DNA Signature Prep Kit (Illumina®) Primer Mix A and B. DNA sequencing was performed with MiSeq FGx (Illumina®) [12]. The genotypes of each sample was found by using a heterozygote threshold, for the autosomal markers (and X chromosome markers in samples belonging to female contributors), of 40% of the maximum coverage of the marker. Furthermore, we used a minimum detection threshold of 10 for the called allele sequenced (stutter sequences could have coverage as low as one).

Allele sequences were excluded from the analysis if they were the parent or stutter of another allele sequence (i.e. if difference between

two heterozygotes was exactly one motif). We also excluded stutter sequences with multiple parental alleles for reasons explained in section 2.3. In total 68,045 stutter sequences were identified of those 2,838 were removed because the allele sequence was a parent or stutter of another allele sequence and 768 were removed because the stutter had multiple identified parents leaving 64,439 observations.

2.2 The block length of the missing motif - BLMM

The definition of block length of the missing motif (BLMM) is based on the same hypothesis underlying allele length and the longest uninterrupted stretch (LUS). Hence, the underlying hypothesis is that the more repetitive a stretch of the strand is, the more likely the PCR process is to stutter. In order to define the BLMM, we first need to define 'a block'.

2.2.1 Block

We define a block of a motif, m, of an allele, A, as an uninterrupted stretch of m in A, such that it is not fully included in another block. We see the blocks of an allele as the identifiable points from which a motif could have stuttered, and the length of the block as having an influence on the probability that the PCR process stutters. Thus, the BLMM is an extension of the LUS concept. The LUS is equivalent to the longest block of the parent allele.

Figure B.1 shows the blocks of the string (also shown in Eq. (1.2)):

$$[\texttt{AATG}]_{10}[\texttt{ACTT}]_4,$$

and will in the remainder of this subsection describe the five blocks found in the string.

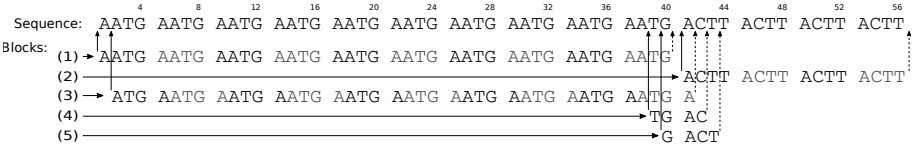

Fig. B.1: A diagram of the blocks contained in the sequence $[\texttt{AATG}]_{10}[\texttt{ACTT}]_4$ and their relation to the sequence. Every vertical solid arrow constitutes the start of a block and the dashed arrow of the same length indicates the end of the block.

The two first blocks seen in the figure are easily obtained:

(1) $[\texttt{AATG}]_{10}$: a block with the motif `AATG`, starting at base one, containing 10 repeats of `AATG`, and

(2) $[\texttt{ACTT}]_4$: a block with the motif `ACTT`, starting at base 41, $(4 \cdot 10 + 1)$, containing four repeats of `ACTT`.

The number of repeats contained in a block is the same as the length of the block. Furthermore, it is indicated by the subscript. Thus, the block $[\texttt{AATG}]_{10}$ has the length 10.

By rewriting the strand $[\texttt{AATG}]_{10}[\texttt{ATTC}]_4$, as

$$\texttt{A}[\texttt{ATGA}]_{10}[\texttt{CTTA}]_3\texttt{CTT},$$

we identify the third block:

(3) $[\texttt{ATGA}]_{10}$: a block with the motif `ATGA`, starting at base two, containing 10 repeats of `ATGA`.

Notice that the stretch $[\texttt{CTTA}]_3$ is by definition not a block, as it is fully contained within the $[\texttt{ACTT}]_4$ block. Thus, if a stutter was created by loosing a `CTTA` motif, it would not be identified. However, it would be identified as a lost `ACTT` motif, because the stretch is fully contained in the $[\texttt{CTTA}]_3$ block. Furthermore, the $[\texttt{ATGA}]_{10}$ block is entirely dependent on the `A` at the 41^{st} base. If the `A` was not there, we would just have a $[\texttt{ATGA}]_9$ stretch, which would be fully included in the $[\texttt{AATG}]_{10}$ block.

Working with compound (or complex) STRs, we will always be able to find short blocks between two directly adjacent blocks. As illustrated in Figure B.1, we found two blocks of length one when moving from the $[\texttt{AATG}]_{10}$ block to the $[\texttt{ACTT}]_4$ block:

(4) $[\texttt{TGAC}]_1$: a block with the motif `TGAC`, starting at base 39, $(4 \cdot 10 - 1)$, containing a repeat of `TGAC`, and

(5) $[\texttt{GACT}]_1$: a block with the motif `GACT`, starting at base 40, $(4 \cdot 10)$, containing a repeat of `GACT`.

Note that because we assume a stutter to be exactly one motif shorter, we are only interested in blocks with motifs of the same length

as the region. In the example above, the region contained tetranucleotides. Thus, we searched for motifs of length four. If the region in question contained tri- or penta-nucleotides, we would have searched for motifs of length three and five, respectively.

Given a stutter, all that remains to be determined is which motif is missing and from which block it originated when comparing the stutter sequence to a potential parental allele.

2.2.2 Determining the missing motif and the BLMM

In order to determine the missing motif and identify the block to find its BLMM, we aligned the potential stutter to a potential parental allele. We defined a stutter as a string missing exactly one motif when compared to its parental allele. In the case of a tetranucleotide, a potential stutter would be missing four consecutive bases when compared to the parental allele.

When aligning the stutter sequence, a, with a potential parental allele, A, the two strings are matched from left to right. When a no longer matches A, a gap is opened and extended. If the number of consecutive extensions is equal to the motif length (in that region), A is classified as a parent of a. It follows that from aligning the stutter sequence with a potential parental allele, we get a missing motif and the block from which it originated. However, because of the way the alignment is performed, it is in some cases only possible to determine the missing motif up to a shift of the motif. As an example, assume that the parental allele, A, is given by the string in Eq. (1.2), and assume that we have observed the following stutter sequence, a:

$$[\texttt{AATG}]_9[\texttt{ACTT}]_4. \tag{B.3}$$

Stutter sequence a could have been created by loosing a motif from one of the following two blocks:

(1) $[\texttt{AATG}]_{10}$ starting at base 1 ending at 40, or

(2) $[\texttt{ATGA}]_{10}$ starting at base 2 ending at 41.

As illustrated in Figure B.2, when aligning from left to right, the stutter would match the parental allele until the 37th base. At this point, the parental allele has an ATGA motif, which did not appear in the stutter

sequence. We would, therefore, identify ATGA as the missing motif in both cases. Thus, the motif can in some cases only be determined upto a shift of the motif.

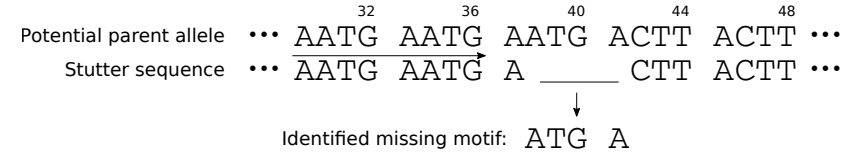

Fig. B.2: The stutter sequence, $[\texttt{AATG}]_9[\texttt{ACTT}]_4$, aligned to a potential parent, $[\texttt{AATG}]_{10}[\texttt{ACTT}]_4$. The two sequences match up until the 37^{th} base, at which point the potential parent has an additional ATGA not found in the stutter sequence.

Given the missing motif and its position, we can find the block from which the missing motif (or a shift thereof) stuttered. The BLMM is the length of this block, thereby, giving it the name: *The block length of the missing motif*.

In the example above, the BLMM would be 10 for both of the stutters. In fact, whenever a missing motif can only be determined upto a shift, all of the shifted blocks from which the motif could have originated would always have the same length. Resulting in the same predictor for our regression model.

In order to showcase the difference between BLMM and LUS, we return to the example used in the introduction. That is, let the parental allele be defined by:

$$[\texttt{AATG}]_{10}[\texttt{ACTT}]_4$$

with two observed stutter sequences defined by:

$$[\texttt{AATG}]_9[\texttt{ACTT}]_4 \text{ and } [\texttt{AATG}]_{10}[\texttt{ACTT}]_3.$$

The LUS of the parental allele is 10. Thus, the LUS assigned to both stutter sequences is 10. However, when the stutter sequences are aligned to the parental allele from left-to-right, we will for the first stutter sequence identify a missing ATGA from the $[\texttt{ATGA}]_{10}$ block and for the second stutter sequence identify a missing ACTT from the $[\texttt{ACTT}]_4$ block. Yielding BLMMs of 10 and 4, respectively. The hypothesis is that the stutter sequences identifying a missing BLMM of 10 would have a higher rate of stuttering than those with a missing BLMM of 4. Therefore, the sequenced stutters of the larger BLMM would result in a larger coverage.

2.3 Stutters with multiple potential parents

Above, we only considered the cases where the potential stutter has a single potential parent. However, if an individual had two different alleles of the same length (i.e. heterozygous in MPS, but homozygous in CE), or if the sample contained DNA from multiple contributors, then a stutter sequence could have multiple potential parental alleles.

To illustrate this point: Assume that we have two alleles A_1 and A_2 with equal lengths and given by:

$$[\texttt{AATG}]_{10}[\texttt{ACTT}]_4 \quad \text{and} \quad [\texttt{AATG}]_9[\texttt{ACTT}]_5.$$

Then the stutter sequence a given by $[\texttt{AATG}]_9[\texttt{ACTT}]_4$ would have received stutter product from both A_1 and A_2, as depicted in Figure B.3. The BLMM of a given A_1 would be 10 while the BLMM of A_2 would be 5. The difference in BLMMs could potentially be large and, thus, the difference in stutter ratio could be large. In fact, the observed coverage of a would have been affected by stuttering from both of the parental alleles. Thus, the coverage would be inflated when compared to stutter sequences with a single parental allele. Therefore, we removed such observations from consideration.

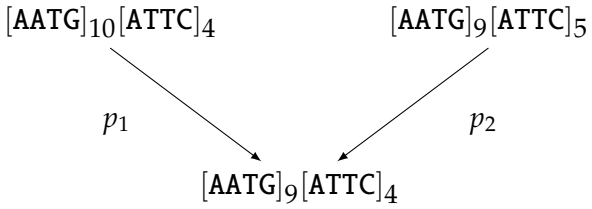

Fig. B.3: Sketch of the two alleles of equal length creating the same stutter product, where p_1 and p_2 are probability of the stutter being created from allele $[\texttt{AATG}]_{10}[\texttt{ACTT}]_4$ and $[\texttt{AATG}]_9[\texttt{ACTT}]_5$, respectively.

2.4 Modelling stutter ratio

We modelled the relationship between the BLMM and the stutter ratio using a linear model with intercept through $(1, 0)$ in accordance with the derivation seen in Appendix A. That is,

$$S_R = \beta \cdot (\text{BLMM} - 1). \tag{B.4}$$

Note by subtracting one from the predictor we have moved the intercept to the origin, which is equivalent to the predictor intercepting the abscissa axis at 1.

As the MPS process suffers from very large marker imbalances, we also examined the effect of marker dependence on the slope. We further note that in forcing the intercept of the model through zero, we avoided the problem of predicting any negative stutter ratios.

Because of the very large amount of data (see Section 2.1), even minute differences in marker specific parameter estimates will be statistically significant. We are interested in assessing the predictive capabilities of the models. Therefore, we compared the models using cross validation (CV). The most common choice of CV is k-fold CV with k equal to 5 or 10, see e.g. [13, Chapter 7]. However, we chose a special case of k-fold CV: the leave-one-out cross validation (LOOCV) method. We chose to use LOOCV, because if we had chosen k-fold CV, then we would have needed to ensure that every motif and marker was included in every training set. This implies that the training and test sets would not have been created randomly and, thus, defeating the purpose of using k-fold CV. The LOOCV method works as follows: (1) use one data point as the validation set, while using the remaining data as a training set, (2) calculate the squared validation error, and (3) repeat (1)-(2) for every data point in the data set. The LOOCV error is then the average squared validation error.

3 Results

3.1 Comparing LUS and BLMM as predictors of stutter ratio

At the left-hand side of Fig. B.4, the stutter ratio is plotted against the LUS. We see large difference in stutter ratios between stutter sequences with the same LUS. Looking more closely at this difference, the larger stutter ratios corresponded to stutter sequences having lost a motif from the longest uninterrupted stretch of the parental allele. While the smaller stutter ratios corresponded to stutter sequences which had lost a motif from a region other than the longest uninterrupted stretch, thereby, splitting the plot into two parts. Comparing the left- and right-hand side of the figure, we see that this split could be explained by the BLMM. This is the case for all markers with compound or complex

motif structures. Similar plots for the remaining markers can be found in the supplementary material.

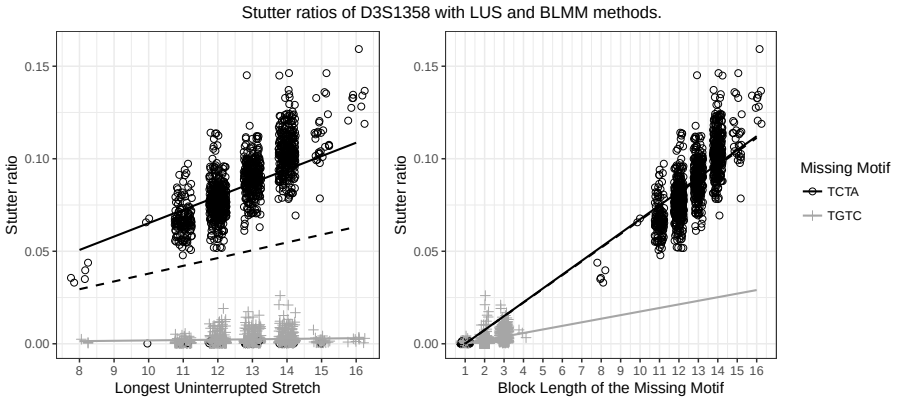

Fig. B.4: Stutter ratios of D13S317 plotted against the longest uninterrupted stretch (LUS) and the block length of the missing motif (BLMM) shown on the left and right, respectively. The points are shaded according to the motifs missing from the stutter sequence when compared to those of the parental allele. The dashed and solid lines correspond to the models with (1) only a marker specific intercept, and (2) marker and motif specific intercepts, respectively. Note that a jitter has been added to visualise the density of the points.

The colouring in Fig. B.4 was made according to the missing motif, when sequences of the stutter sequence were compared to that of the parental allele. We see that the missing motif could explain the split seen in the LUS. Fig. B.5 shows a similar split when using BLMM as a predictor, but that split is much less pronounced when compared to that of the LUS.

Thus, we also considered models of the form:

$$S_R = \beta_{\text{Missing motif}} \cdot x,$$

where x is either $(\text{BLMM} - 1)$ or $(\text{LUS} - 1)$. The slope now depends on the missing motif. Furthermore, we also considered a marker dependent version of this model, i.e. $S_R = \beta_{\text{Missing motif, Marker}} \cdot x$.

Table B.1 shows the LOOCV errors for each of the four models using both LUS and BLMM as predictors. We see that the model with one common slope for all markers had by far the largest error, and that the model with marker and motif dependent slopes had the smallest.

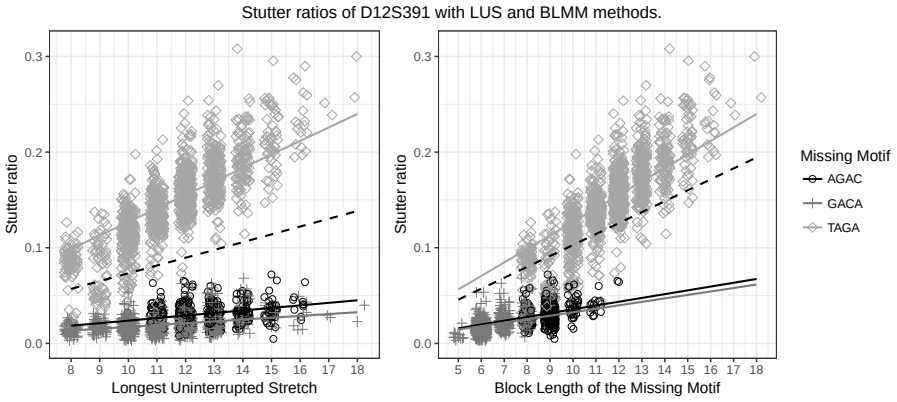

Fig. B.5: Stutter ratios of D12S391 plotted against the longest uninterrupted stretch (LUS) and the block length of the missing motif (BLMM) shown on the left and right, respectively. The points are shaded according to the motifs missing from the stutter sequence when compared to those of the parental allele. The dashed and solid lines correspond to the models with (1) only a marker specific intercept, and (2) marker and motif specific intercepts, respectively. Note that a jitter has been added to visualise the density of the points.

That was expected as these are the least and most flexible models, respectively.

Table B.1: The leave-one-out cross validation (LOOCV) error linear models with the slope dependent on nothing, the missing motif, the marker, and both missing motif and marker, using BLMM and LUS as predictors.

Model	LOOCV		Odds (LUS / BLMM)
	BLMM	LUS	
$S_R = \beta \cdot x$	1.54×10^{-3}	2.99×10^{-3}	1.94
$S_R = \beta_{\text{Missing motif}} \cdot x$	1.02×10^{-3}	1.95×10^{-3}	1.91
$S_R = \beta_{\text{Marker}} \cdot x$	7.42×10^{-4}	2.55×10^{-3}	3.44
$S_R = \beta_{\text{Missing motif, Marker}} \cdot x$	5.20×10^{-4}	1.12×10^{-3}	2.15

Note that the LOOCV was only averaged across the markers with compound and complex repeat structures.

Furthermore, when creating the LOOCV, we restricted the analysis to the compound and complex markers, because there is no difference between the LUS and BLMM for markers with simple repeat struc-

tures. We defined the marker as simple if the LUS was equal to the allele length, and, therefore, excluded the following markers: CSF1PO, D5S818, D7S820, D10S1248, D13S317, D17S1301, D18S51, D20S482, PENTAD, PENTAE, TPOX, DXS8378, HPRTB, DYS391, DYS438, DYS439, DYS460, DYS549, DYS576, and DYS643.

In all four cases, the BLMM models had a smaller LOOCV than the LUS models. Furthermore, we noticed that the LOOCV of the BLMM model dependent on marker was slightly lower than that of the LUS model dependent on both marker and the missing motif.

Supplementary Table B.2 shows the LOOCV for every marker (including markers with simple repeat structures) of the two marker dependent models. We see that for markers containing compound or complex repeat structures (e.g. D3S1358, FGA, vWA, etc.) the BLMM did better than the LUS in every single case. Furthermore, we see that the difference in LOOCV can be rather large. Looking at the model not dependent on the missing motif, we see that the LOOCV decreased by a factor upto 17.5, 16.5, 17 for the autosomal, X, and Y STRs, respectively, when switching from LUS to BLMM.

Table B.2: The leave-one-out cross validation (LOOCV) error shown of 27 autosomal, 7 X, and 23 Y markers for the marker dependent models presented above using both predictors (BLMM and LUS).

		LOOCV					
		$S_R = \beta_{\text{Missing motif, Marker}} \cdot x$			$S_R = \beta_{\text{Marker}} \cdot x$		
Marker	Type	BLMM	LUS	LUS/BLMM	BLMM	LUS	LUS/BLMM
Autosomal	CSF1PO	1.74×10^{-4}	1.74×10^{-4}	1.00	1.74×10^{-4}	1.74×10^{-4}	1.00
Autosomal	D5S818	4.81×10^{-4}	4.81×10^{-4}	1.00	4.81×10^{-4}	4.81×10^{-4}	1.00
Autosomal	D7S820	2.07×10^{-4}	2.07×10^{-4}	1.00	2.07×10^{-4}	2.07×10^{-4}	1.00
Autosomal	D10S1248	2.78×10^{-4}	2.78×10^{-4}	1.00	2.78×10^{-4}	2.78×10^{-4}	1.00
Autosomal	D13S317	1.09×10^{-4}	1.09×10^{-4}	1.00	1.09×10^{-4}	1.09×10^{-4}	1.00
Autosomal	D17S1301	4.83×10^{-4}	4.83×10^{-4}	1.00	4.83×10^{-4}	4.83×10^{-4}	1.00
Autosomal	D18S51	2.77×10^{-4}	2.77×10^{-4}	1.00	2.77×10^{-4}	2.77×10^{-4}	1.00
Autosomal	D20S482	1.73×10^{-4}	1.73×10^{-4}	1.00	1.73×10^{-4}	1.73×10^{-4}	1.00
Autosomal	PENTAD	6.01×10^{-5}	6.01×10^{-5}	1.00	6.01×10^{-5}	6.01×10^{-5}	1.00
Autosomal	PENTAE	1.70×10^{-4}	1.70×10^{-4}	1.00	1.70×10^{-4}	1.70×10^{-4}	1.00
Autosomal	TPOX	7.46×10^{-5}	7.46×10^{-5}	1.00	7.46×10^{-5}	7.46×10^{-5}	1.00
Autosomal	D1S1656	1.77×10^{-3}	2.51×10^{-3}	1.42	1.79×10^{-3}	2.97×10^{-3}	1.66
Autosomal	D2S1338	3.19×10^{-4}	5.77×10^{-4}	1.81	1.23×10^{-3}	3.65×10^{-3}	2.96
Autosomal	D2S441	6.88×10^{-5}	7.80×10^{-5}	1.13	6.88×10^{-5}	1.45×10^{-4}	2.11
Autosomal	D3S1358	8.54×10^{-5}	2.22×10^{-4}	2.60	1.11×10^{-4}	1.94×10^{-3}	17.5
Autosomal	D4S2408	8.99×10^{-5}	8.99×10^{-5}	9.99×10^{-1}	8.98×10^{-5}	9.62×10^{-5}	1.07
Autosomal	D6S1043	9.54×10^{-5}	9.54×10^{-5}	1.00	2.35×10^{-4}	1.39×10^{-3}	5.91
Autosomal	D8S1179	3.32×10^{-4}	1.14×10^{-3}	3.45	3.33×10^{-4}	4.23×10^{-3}	12.70

Autosomal	D9S1122	2.40×10^{-4}	2.43×10^{-4}	1.01	2.40×10^{-4}	1.17×10^{-3}	4.86
Autosomal	D12S391	4.26×10^{-4}	4.33×10^{-4}	1.02	2.02×10^{-3}	4.54×10^{-3}	2.25
Autosomal	D16S539	2.01×10^{-4}	2.03×10^{-4}	1.01	2.01×10^{-4}	2.03×10^{-4}	1.01
Autosomal	D19S433	1.75×10^{-4}	1.78×10^{-4}	1.02	1.75×10^{-4}	1.16×10^{-3}	6.63
Autosomal	D21S11	2.52×10^{-4}	6.93×10^{-4}	2.75	4.90×10^{-4}	8.57×10^{-4}	1.75
Autosomal	D22S1045	6.73×10^{-4}	6.73×10^{-4}	1.00	6.75×10^{-4}	1.42×10^{-3}	2.11
Autosomal	FGA	3.70×10^{-4}	2.27×10^{-3}	6.13	3.80×10^{-4}	3.86×10^{-3}	10.20
Autosomal	TH01	8.68×10^{-5}	8.68×10^{-5}	1.00	1.97×10^{-4}	5.83×10^{-4}	2.96
Autosomal	VWA	6.46×10^{-4}	7.44×10^{-4}	1.15	6.80×10^{-4}	2.82×10^{-3}	4.14
X	DXS8378	4.63×10^{-5}	4.63×10^{-5}	1.00	4.63×10^{-5}	4.63×10^{-5}	1.00
X	HPRTB	8.37×10^{-5}	8.37×10^{-5}	1.00	8.37×10^{-5}	8.37×10^{-5}	1.00
X	DXS7132	1.91×10^{-4}	1.92×10^{-4}	1.01	1.91×10^{-4}	1.92×10^{-4}	1.01
X	DXS7423	4.05×10^{-5}	3.42×10^{-4}	8.45	4.05×10^{-5}	3.86×10^{-4}	9.53
X	DXS10074	9.58×10^{-5}	1.51×10^{-4}	1.58	9.65×10^{-5}	7.58×10^{-4}	7.85
X	DXS10103	1.94×10^{-3}	1.95×10^{-3}	1.00	1.96×10^{-3}	2.54×10^{-3}	1.30
X	DXS10135	2.05×10^{-4}	1.78×10^{-3}	8.66	2.36×10^{-4}	3.91×10^{-3}	16.50
Y	DYS391	1.90×10^{-4}	1.90×10^{-4}	1.00	1.90×10^{-4}	1.90×10^{-4}	1.00
Y	DYS438	1.57×10^{-5}	1.57×10^{-5}	1.00	1.57×10^{-5}	1.57×10^{-5}	1.00
Y	DYS439	1.50×10^{-4}	1.50×10^{-4}	1.00	1.50×10^{-4}	1.50×10^{-4}	1.00
Y	DYS460	2.73×10^{-4}	2.73×10^{-4}	1.00	2.73×10^{-4}	2.73×10^{-4}	1.00
Y	DYS549	7.73×10^{-5}	7.73×10^{-5}	1.00	7.73×10^{-5}	7.73×10^{-5}	1.00
Y	DYS576	1.25×10^{-4}	1.25×10^{-4}	1.00	1.25×10^{-4}	1.25×10^{-4}	1.00
Y	DYS643	3.73×10^{-5}	3.73×10^{-5}	1.00	3.73×10^{-5}	3.73×10^{-5}	1.00
Y	DYF387S1	5.15×10^{-3}	5.55×10^{-3}	1.08	5.72×10^{-3}	7.58×10^{-3}	1.32
Y	DYS19	2.11×10^{-4}	1.09×10^{-3}	5.18	2.11×10^{-4}	1.43×10^{-3}	6.79
Y	DYS385	1.28×10^{-4}	6.71×10^{-4}	5.26	8.71×10^{-4}	3.48×10^{-3}	3.99
Y	DYS390	1.77×10^{-4}	8.09×10^{-4}	4.56	3.33×10^{-4}	9.37×10^{-4}	2.81
Y	DYS392	6.32×10^{-3}	6.48×10^{-3}	1.03	6.32×10^{-3}	6.48×10^{-3}	1.03
Y	DYS437	2.70×10^{-4}	1.16×10^{-3}	4.30	2.76×10^{-4}	1.29×10^{-3}	4.70
Y	DYS448	1.38×10^{-5}	1.39×10^{-5}	1.01	1.41×10^{-5}	2.38×10^{-5}	1.69
Y	DYS461	1.80×10^{-3}	1.85×10^{-3}	1.03	1.80×10^{-3}	2.51×10^{-3}	1.39
Y	DYS481	1.54×10^{-3}	1.54×10^{-3}	1.00	1.55×10^{-3}	1.06×10^{-2}	6.84
Y	DYS505	6.88×10^{-5}	6.85×10^{-5}	9.96×10^{-1}	6.87×10^{-5}	7.22×10^{-4}	10.50
Y	DYS522	1.50×10^{-4}	1.54×10^{-4}	1.03	1.50×10^{-4}	1.54×10^{-4}	1.03
Y	DYS533	1.68×10^{-4}	1.68×10^{-4}	1.00	1.68×10^{-4}	2.69×10^{-4}	1.61
Y	DYS570	2.73×10^{-4}	3.91×10^{-4}	1.43	2.73×10^{-4}	3.91×10^{-4}	1.43
Y	DYS612	5.77×10^{-4}	1.18×10^{-2}	20.4	1.06×10^{-3}	1.80×10^{-2}	17.0
Y	DYS635	1.76×10^{-4}	7.26×10^{-4}	4.12	1.77×10^{-4}	8.31×10^{-4}	4.69
Y	Y-GATA-H4	4.95×10^{-4}	4.93×10^{-4}	9.94×10^{-1}	4.95×10^{-4}	2.24×10^{-3}	4.52

4 Discussion

The strength of moving from LUS to BLMM is clearly seen when visually comparing the two predictors plotted against the stutter ratio in Fig. B.4. This point is further emphasised, in Table B.1. It is clear that BLMM is a better predictor of stutter ratio than LUS. It is also clear that it is beneficial to make the model dependent on both marker and motif. However, including the missing motif as a predictor may not be a viable option unless we are absolutely sure that the sample used to

fit the parameters is representative of the population. If it is not, we might find a motif in the population not represented in the sample. In that case, we would not be able to predict the stutter ratio making it a poor choice of model for predictive purposes. Therefore, we advocate to use the marker dependent model, even though we in Fig. B.5 saw that there is a small difference between the motifs for the BLMM.

Furthermore, the BLMM avoids the need for a two component mixture model as the one presented in Bright et al. (2013) [9]. They noted that the second component, with larger variance, hypothetically caught *"those alleles with complex repeat structures comprising variant regions with differing LUS values"*. The BLMM handles this problem.

The supplementary figures show some unexpected patterns, besides those caused by the difference in the stuttering motif. The right-hand side of Fig. B.9 shows the stutter ratio and the BLMM of the D2S441 marker. The stutter ratio for a BLMM of 8 was significantly larger than the expected stutter ratio (a similar pattern can be seen in Fig. B.30). We put forward the hypothesis, that this is caused by the repeat composition of the DNA. Two of the most frequent STR structures of the D2S441 marker are: (1) $[\texttt{TCTA}]_x[\texttt{TCTG}]_1[\texttt{TCTA}]_1$ and (2) $[\texttt{TCTA}]_x$. Focusing on the stutter sequences with a BLMM of 8, their observed stutter ratios have more in common with the expected stutter ratios of stutter sequences with larger BLMM's. Furthermore, the repeat structure of the stutter sequences with a BLMM of 8 are dominated by the structure depicted in (1). Our hypothesis is that the PCR process reacts to the $[\texttt{TCTG}]_1$ as if it were a $[\texttt{TCTA}]_1$. That is, the PCR process reacts to the repeat structure $[\texttt{TCTA}]_x[\texttt{TCTG}]_1[\texttt{TCTA}]_1$ as if it were $[\texttt{TCTA}]_{x+2}$, increasing the length of the longest block from x to $x+2$, thereby increasing its rate of stuttering. However, we do not have enough data on the phenomenon to draw any conclusion regarding this hypothesis.

A way of handling this phenomenon was presented in a recent paper Woerner et al. (2018) [14]. They suggest modelling these sequences independently of the regular variants. However, extending the BLMM concept to identify and account for these sequences is not straight forward and, therefore, left to future research.

Trying to utilise more sub-repeat information than just the LUS is not new to MPS. Taylor et al. (2016) [15] defined the multi-sequence model (MSM) as the average length of the blocks most likely to stutter.

The MSM was designed to handle the problem described in Bright et al. (2013), i.e. it would eliminate the need for a two component model. However, the MSM would always be equal for every stutter sequence of the same parental allele. Thus, it would exhibit similar behaviour as that seen of the LUS in Fig. B.4.

Accounting for stuttering can help determine the relationship between the relative contributions of DNA from the contributors in the case of DNA mixtures. Thus, we expect that the primary use of the BLMM will be in modelling the allele coverage in MPS STR DNA mixture samples. Furthermore, it gives some insight into a long considered hypothesis: that the rate of stuttering during PCR amplification increases the more repetitive the DNA strand is.

A The relationship between BLMM and stutter ratio

This appendix aims to determine the relationship between the block length of the missing motif and the stutter ratio. We will assume that we are in the copying stage of the PCR process. Without loss of generality, we can assume that the STR regions are simple, as the probability of stuttering 'resets' when the motif being copied is changed. For simplicity, we will assume (1) that the probability of a repeated motif stuttering is constant and equal to p (the probability of not stuttering is $(1-p)$), (2) that there are only two strings, i.e. the parent allele and the stutter sequence, and (3) the strings have been copied correctly upto and including the r'th repeated motif.

Assume that the probability of a repeated motif stuttering is constant and equal to p (i.e. the probability of not stuttering is $(1-p)$), and that we have correctly copied the parental allele and stutter sequence upto, and including, the r'th repeated motif.

It follows, that we can describe the number of parental allele strands at the $(r+1)$'th motif, $A(r+1)$, as: The number of parental allele strands correctly copied upto the r'th repeat, $A(r)$, multiplied by the probability of not stuttering, $1-p$, implying the equation:

$$A(r+1) = (1-p)A(r). \tag{B.5}$$

Whereas we can describe the number of stutter strands at the $(r+$

1)'th motif, $a(r+1)$ as: The number of stutter strands correctly copied until the r'th repeat, $a(r)$, multiplied by the probability of not stuttering, $1-p$, plus the number of parental allele strands, at the r'th repeat, $A(r)$, times the probability of stuttering, p, which implies the equation:

$$a(r+1) = (1-p)a(r) + pA(r). \tag{B.6}$$

Solving the system of equations defined by Equation (B.5) and (B.6) yields the following solutions:

$$A(r+1) = (1-p)^r A(0) \quad \text{and} \quad a(r+1) = rp(1-p)^{r-1}A(0), \tag{B.7}$$

where $A(0)$ is the number of parental allele strands at the beginning of the process. We assumed that there was no stutter at the beginning of the process, i.e. $a(0) = 0$.

Taking the ratio between the number of stutter and parental strands yields:

$$\frac{a(r+1)}{A(r+1)} = \frac{p}{1-p}r, \tag{B.8}$$

i.e. the relationship between the stutter ratio, $a(r+1)/A(r+1)$, and the repeat position in the strand, r, is linear through the point $(1,0)$. Note that the intercept is not through the origin as the smallest possible value of r is 1, as the parent allele would need at least one motif to lose.

B Supplementary figures

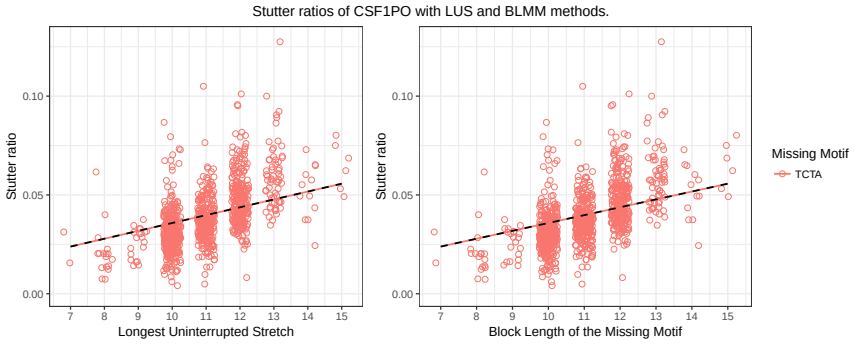

Fig. B.6: The stutter ratio plotted against the LUS and BLMM of CSF1PO. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

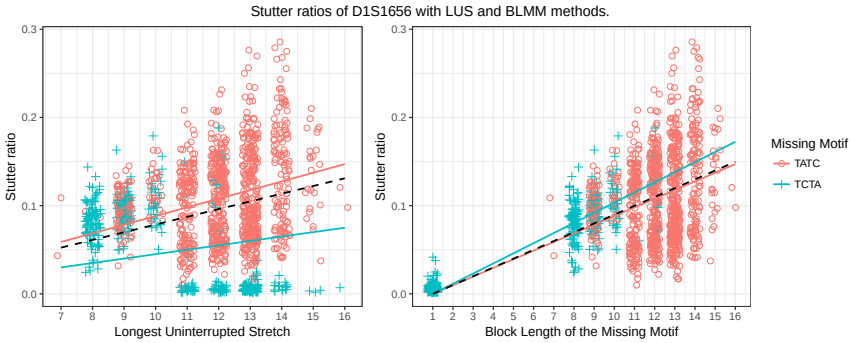

Fig. B.7: The stutter ratio plotted against the LUS and BLMM of D1S1656. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

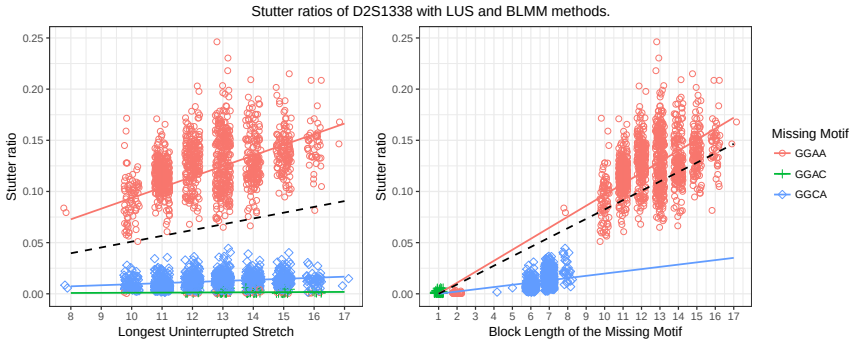

Fig. B.8: The stutter ratio plotted against the LUS and BLMM of D2S1338. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

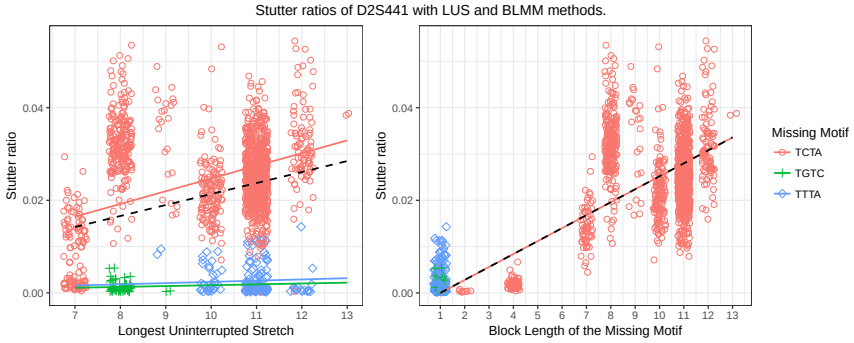

Fig. B.9: The stutter ratio plotted against the LUS and BLMM of D2S441. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

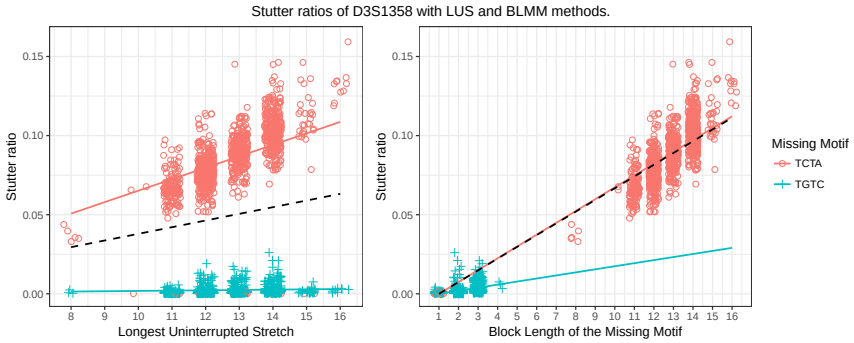

Fig. B.10: The stutter ratio plotted against the LUS and BLMM of D3S1358. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

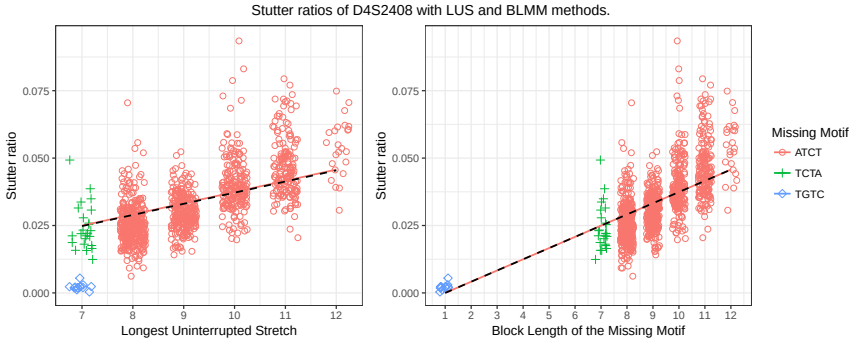

Fig. B.11: The stutter ratio plotted against the LUS and BLMM of D4S2408. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

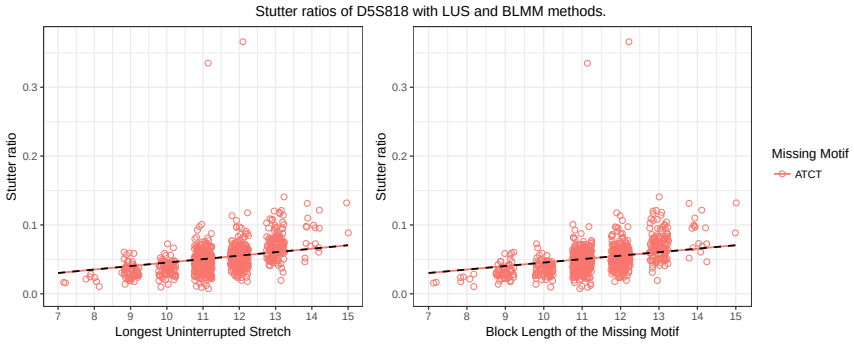

Fig. B.12: The stutter ratio plotted against the LUS and BLMM of D5S818. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

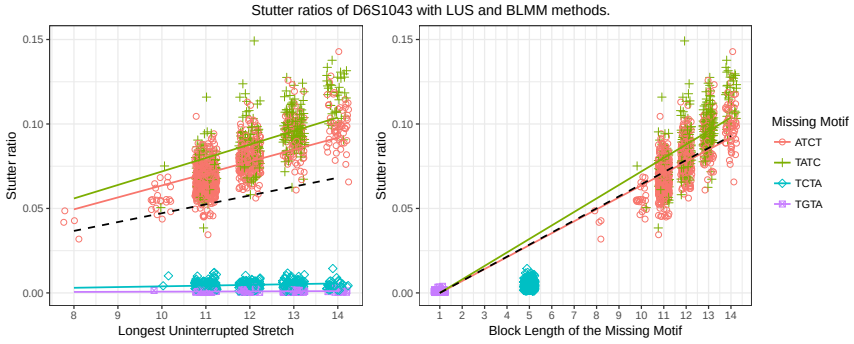

Fig. B.13: The stutter ratio plotted against the LUS and BLMM of D6S1043. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

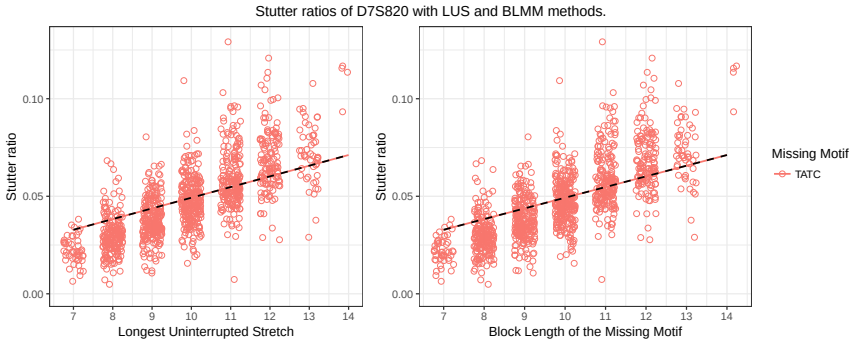

Fig. B.14: The stutter ratio plotted against the LUS and BLMM of D7S820. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

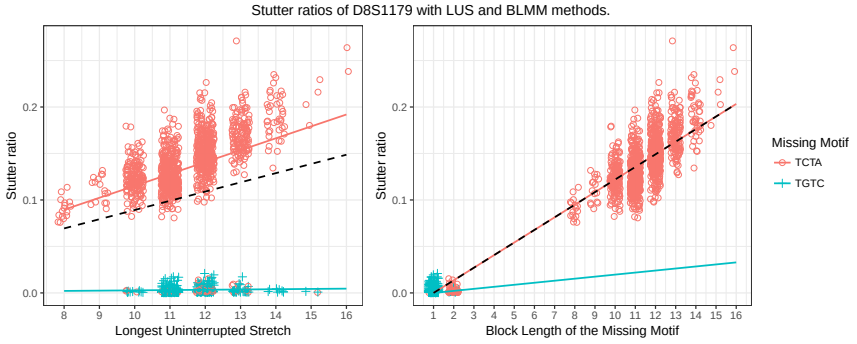

Fig. B.15: The stutter ratio plotted against the LUS and BLMM of D8S1179. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

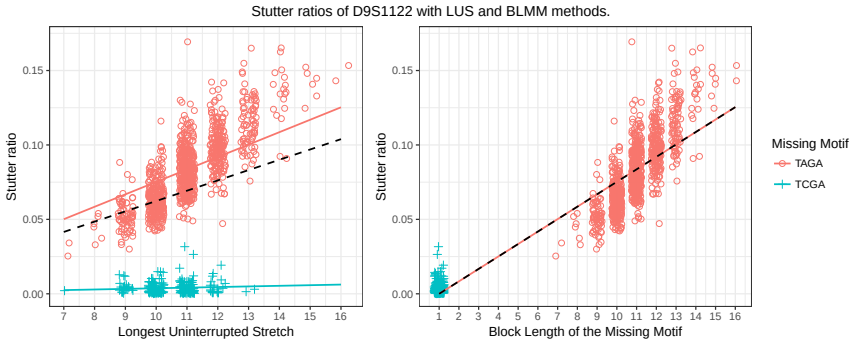

Fig. B.16: The stutter ratio plotted against the LUS and BLMM of D9S1122. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

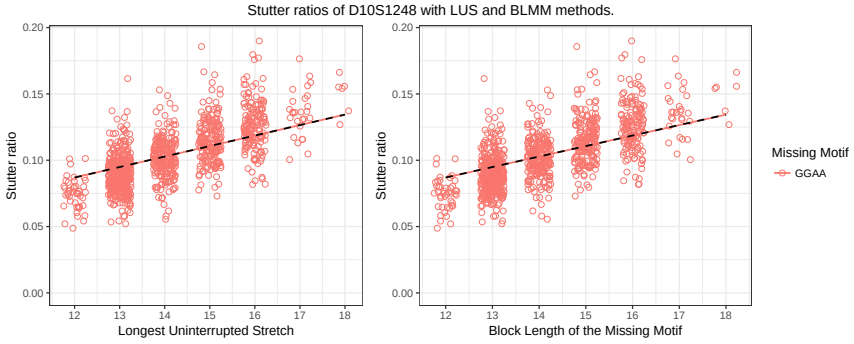

Fig. B.17: The stutter ratio plotted against the LUS and BLMM of D10S1248. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

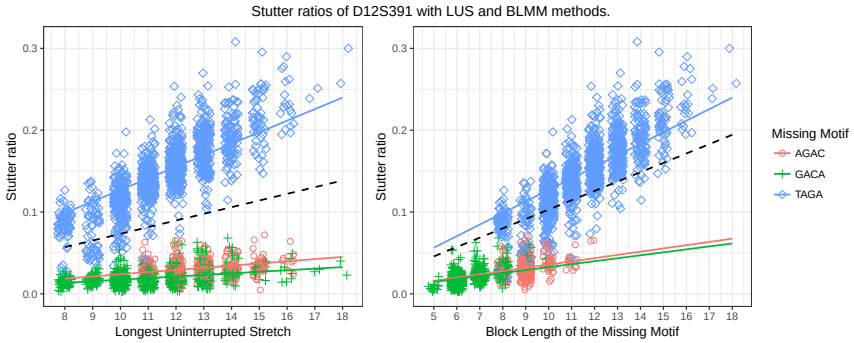

Fig. B.18: The stutter ratio plotted against the LUS and BLMM of D12S391. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

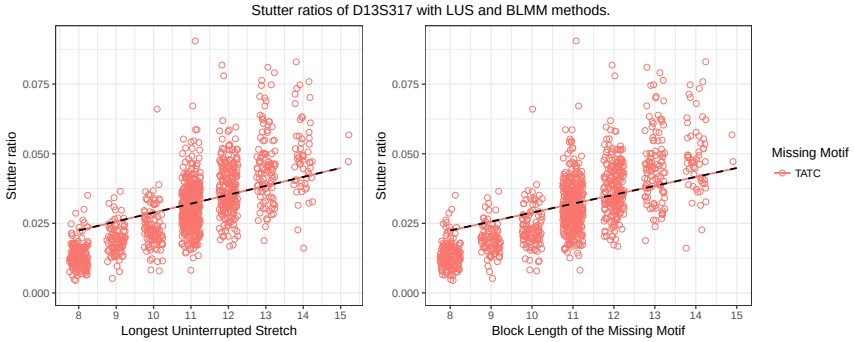

Fig. B.19: The stutter ratio plotted against the LUS and BLMM of D13S317. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

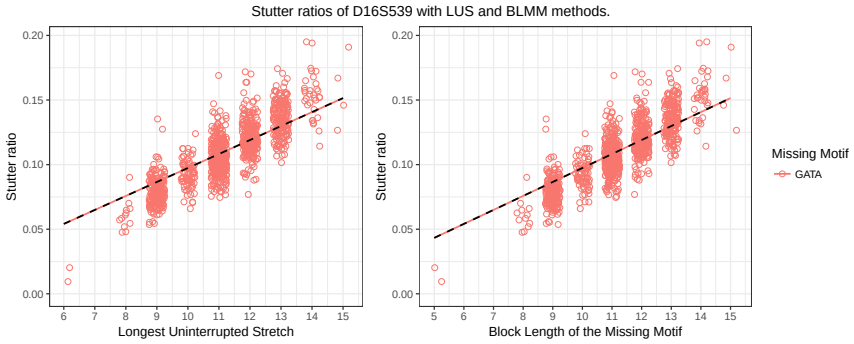

Fig. B.20: The stutter ratio plotted against the LUS and BLMM of D16S539. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

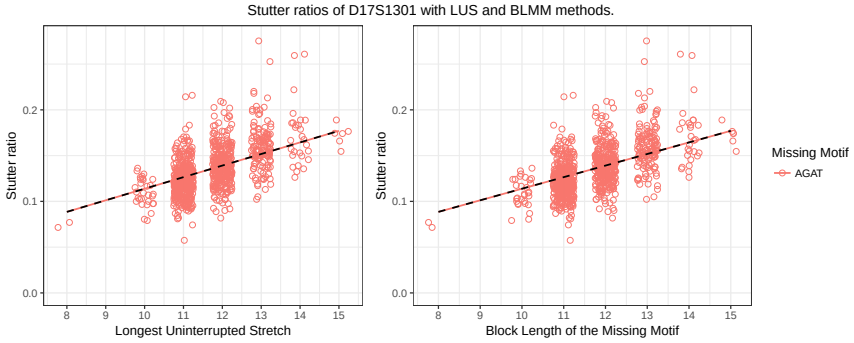

Fig. B.21: The stutter ratio plotted against the LUS and BLMM of D17S1301. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

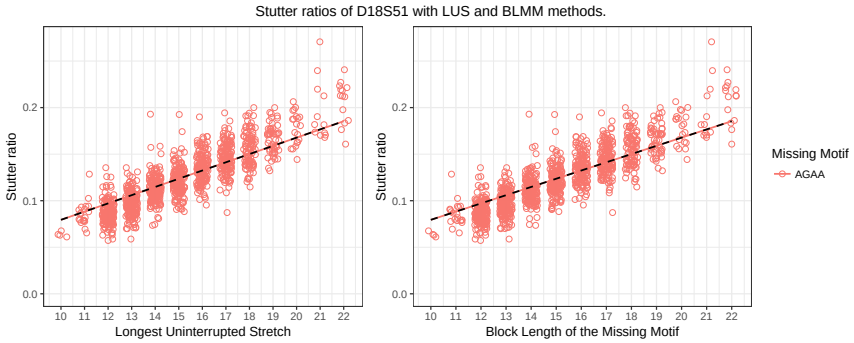

Fig. B.22: The stutter ratio plotted against the LUS and BLMM of D18S51. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

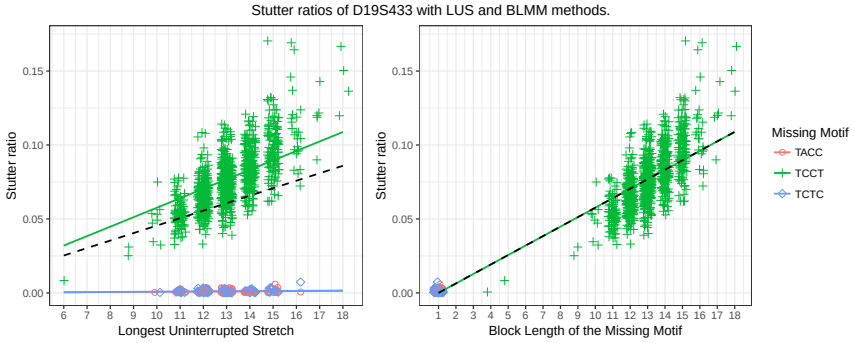

Fig. B.23: The stutter ratio plotted against the LUS and BLMM of D19S433. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

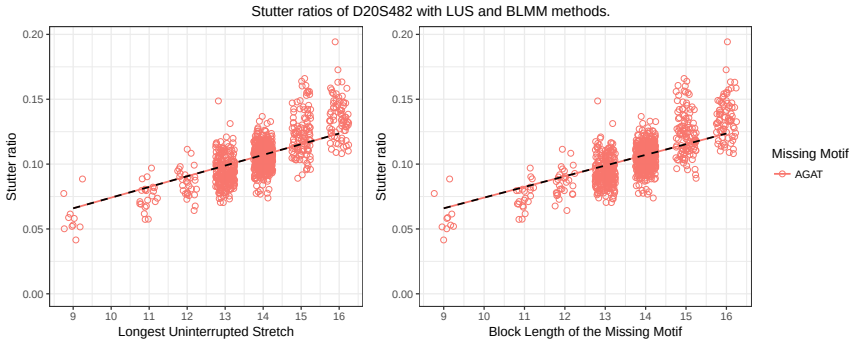

Fig. B.24: The stutter ratio plotted against the LUS and BLMM of D20S482. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

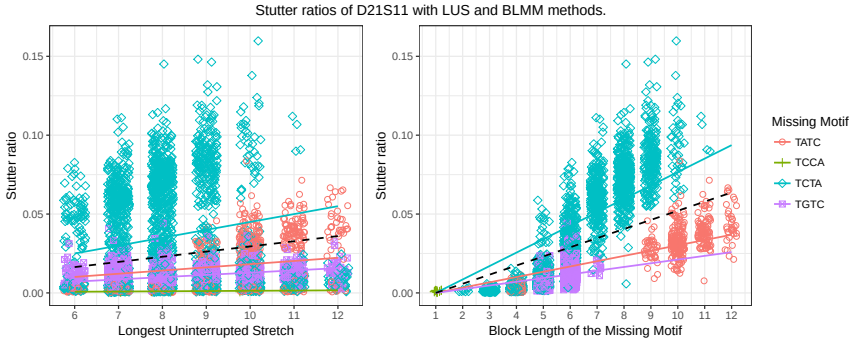

Fig. B.25: The stutter ratio plotted against the LUS and BLMM of D21S11. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

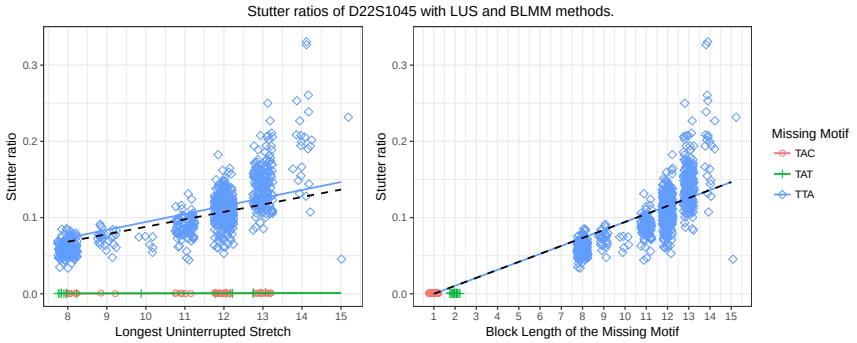

Fig. B.26: The stutter ratio plotted against the LUS and BLMM of D22S1045. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

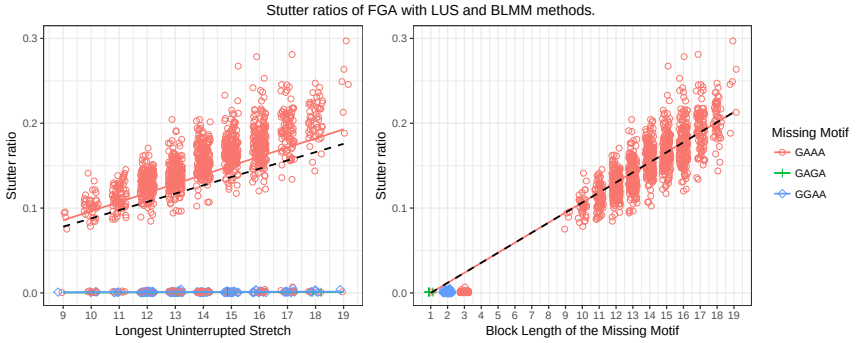

Fig. B.27: The stutter ratio plotted against the LUS and BLMM of FGA. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

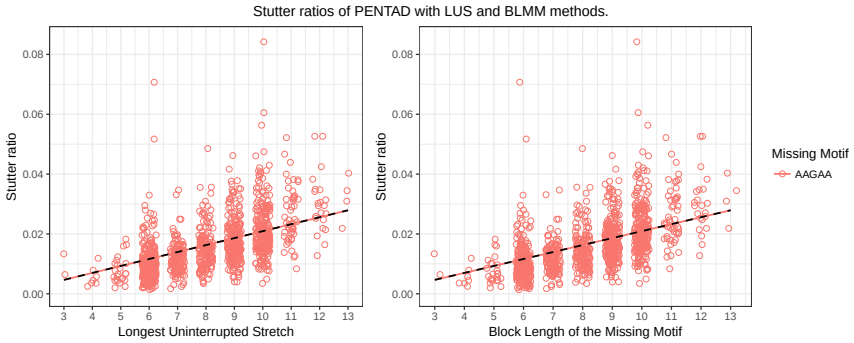

Fig. B.28: The stutter ratio plotted against the LUS and BLMM of PENTAD. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

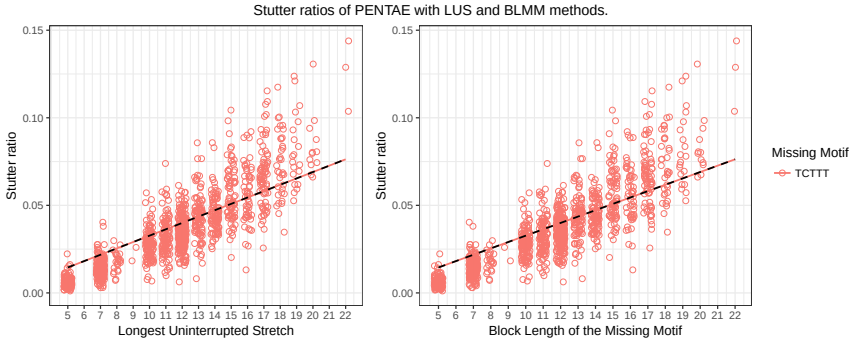

Fig. B.29: The stutter ratio plotted against the LUS and BLMM of PENTAE. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

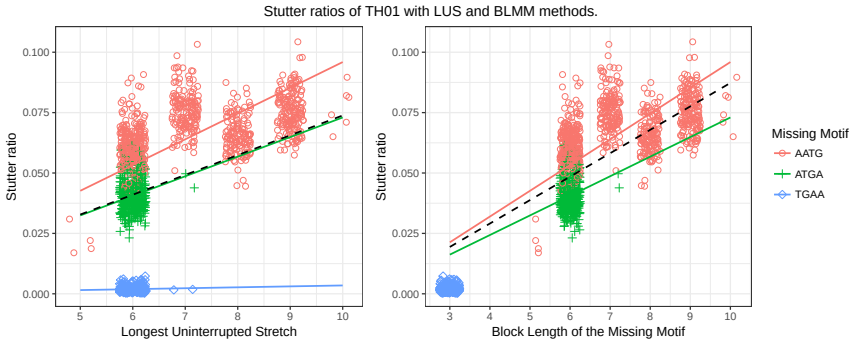

Fig. B.30: The stutter ratio plotted against the LUS and BLMM of TH01. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

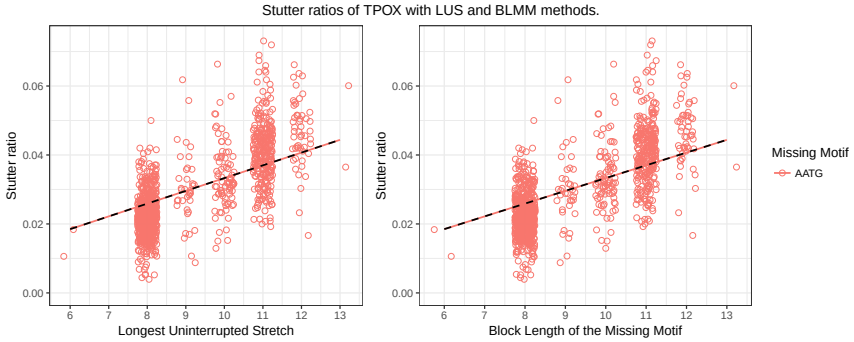

Fig. B.31: The stutter ratio plotted against the LUS and BLMM of TPOX. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

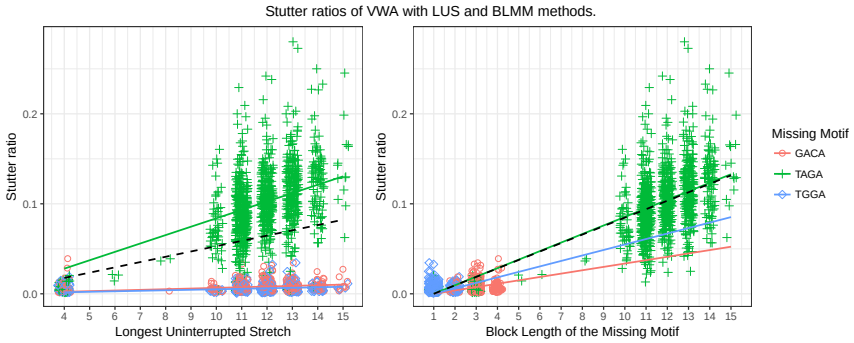

Fig. B.32: The stutter ratio plotted against the LUS and BLMM of VWA. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

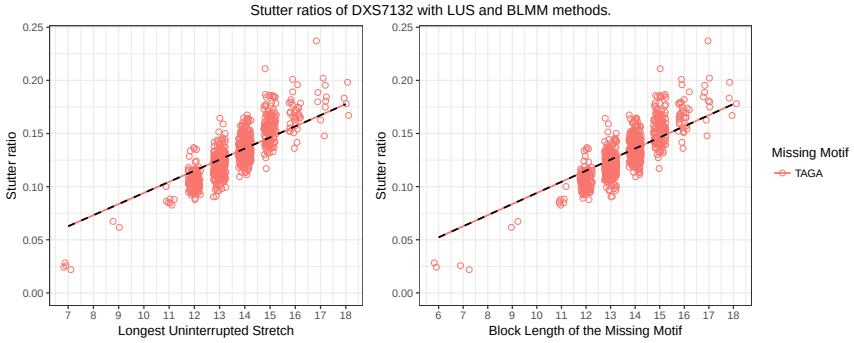

Fig. B.33: The stutter ratio plotted against the LUS and BLMM of DXS7132. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

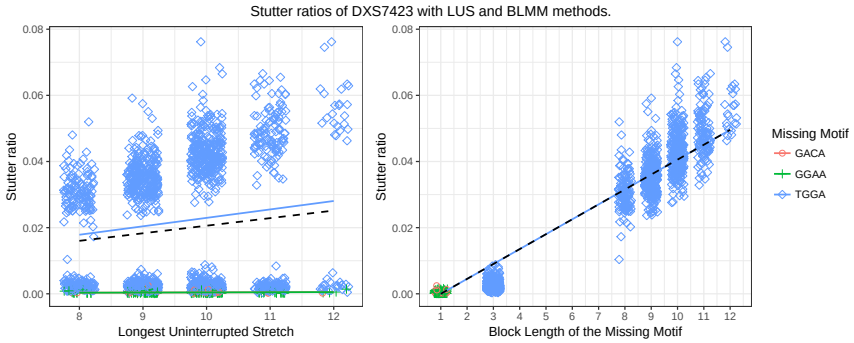

Fig. B.34: The stutter ratio plotted against the LUS and BLMM of DXS7423. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

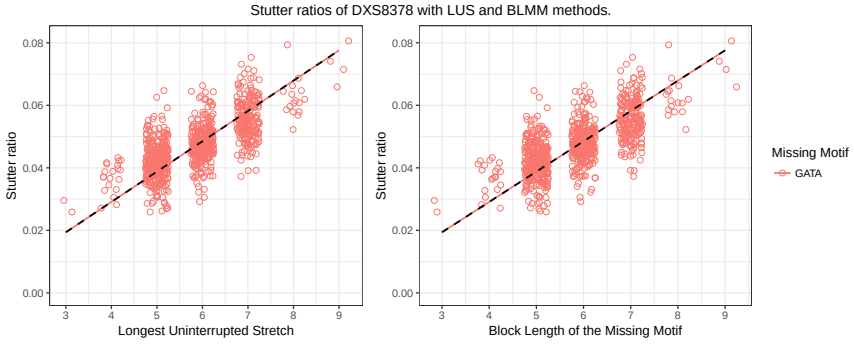

Fig. B.35: The stutter ratio plotted against the LUS and BLMM of DXS8378. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

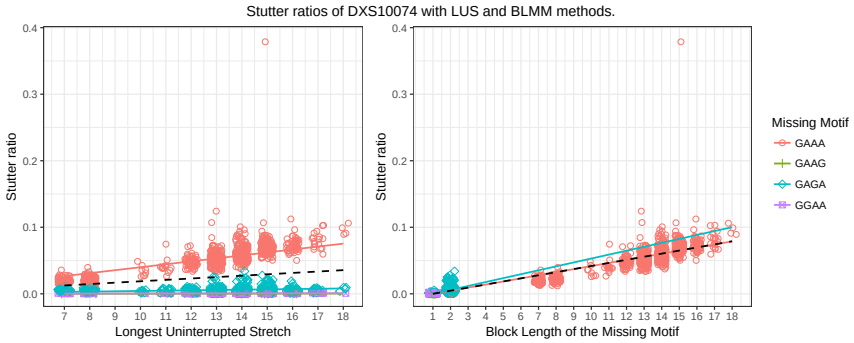

Fig. B.36: The stutter ratio plotted against the LUS and BLMM of DXS10074. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

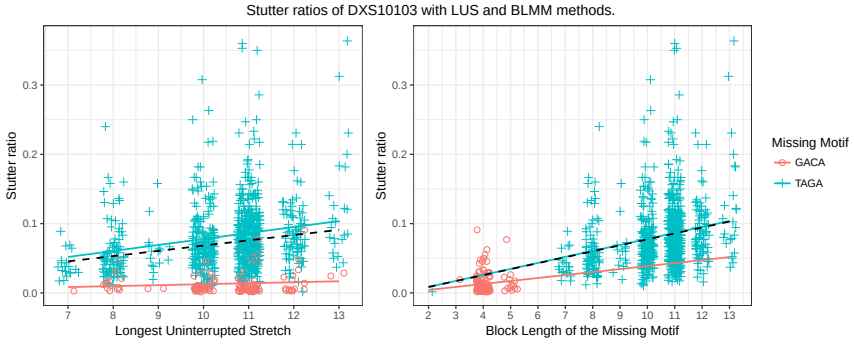

Fig. B.37: The stutter ratio plotted against the LUS and BLMM of DXS10103. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

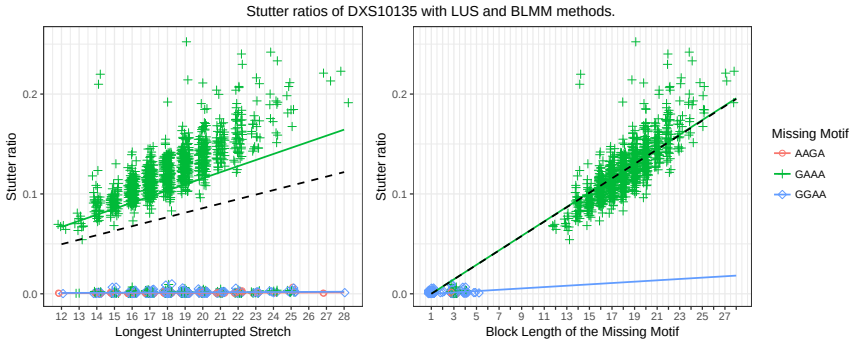

Fig. B.38: The stutter ratio plotted against the LUS and BLMM of DXS10135. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

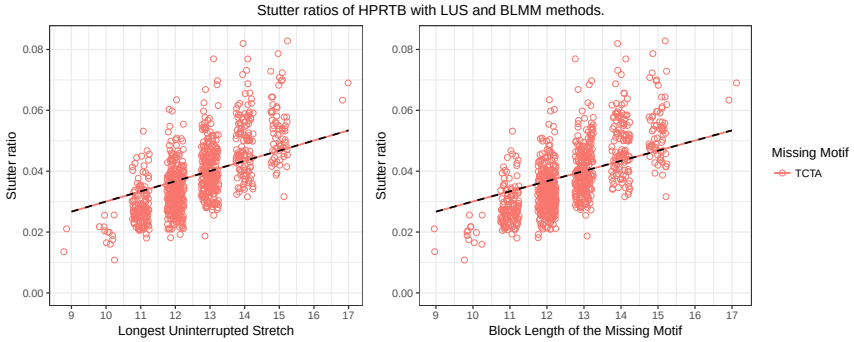

Fig. B.39: The stutter ratio plotted against the LUS and BLMM of HPRTB. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

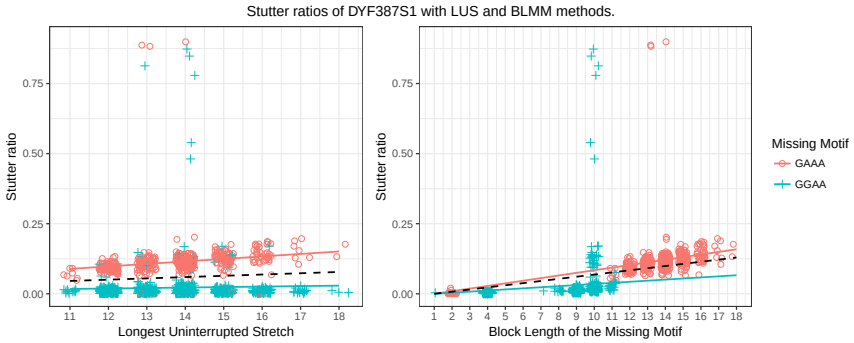

Fig. B.40: The stutter ratio plotted against the LUS and BLMM of DYF387S1. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

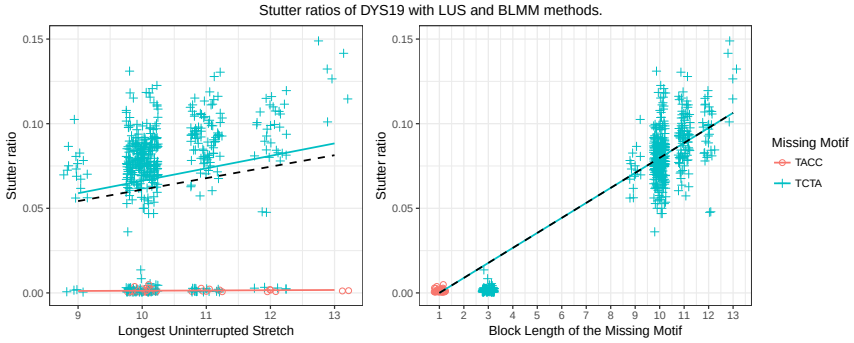

Fig. B.41: The stutter ratio plotted against the LUS and BLMM of DYS19. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

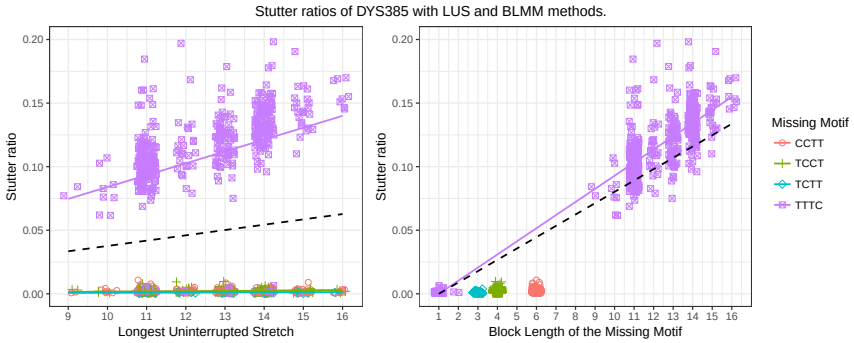

Fig. B.42: The stutter ratio plotted against the LUS and BLMM of DYS385. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

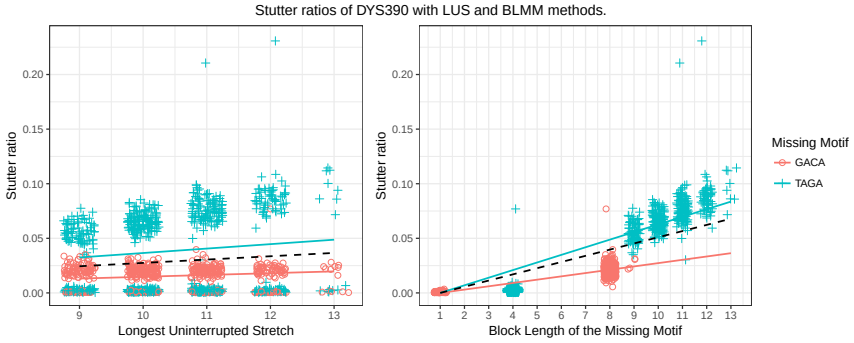

Fig. B.43: The stutter ratio plotted against the LUS and BLMM of DYS390. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

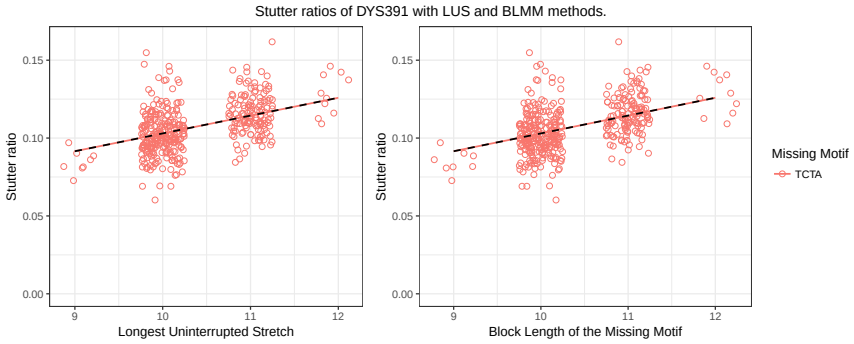

Fig. B.44: The stutter ratio plotted against the LUS and BLMM of DYS391. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

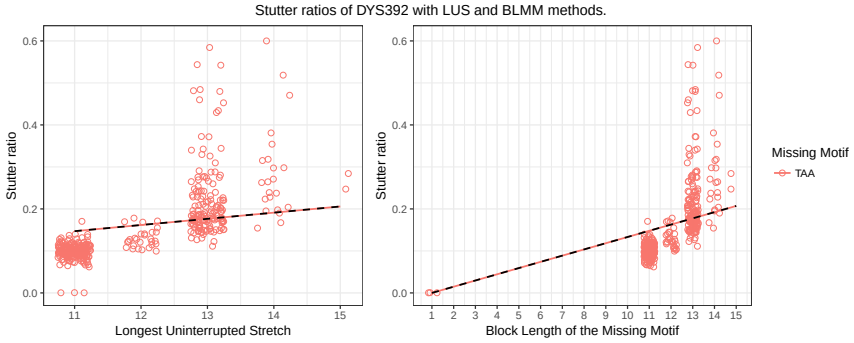

Fig. B.45: The stutter ratio plotted against the LUS and BLMM of DYS392. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

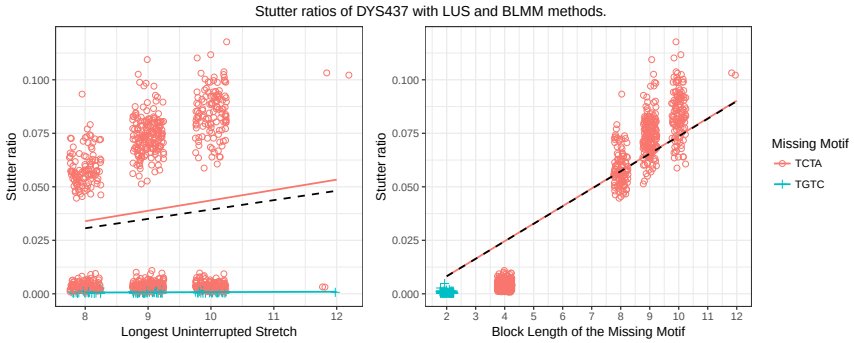

Fig. B.46: The stutter ratio plotted against the LUS and BLMM of DYS437. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

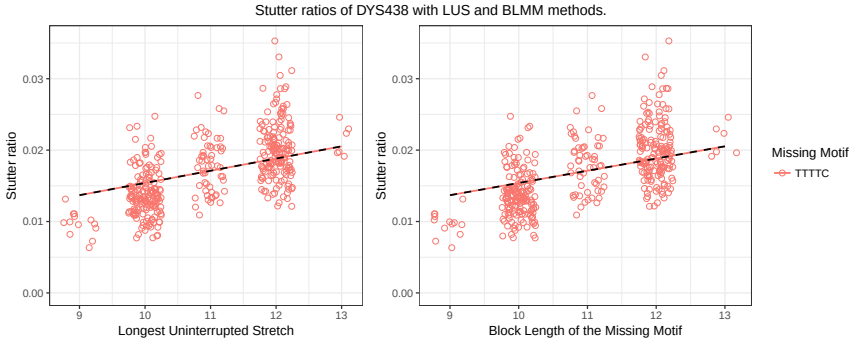

Fig. B.47: The stutter ratio plotted against the LUS and BLMM of DYS438. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

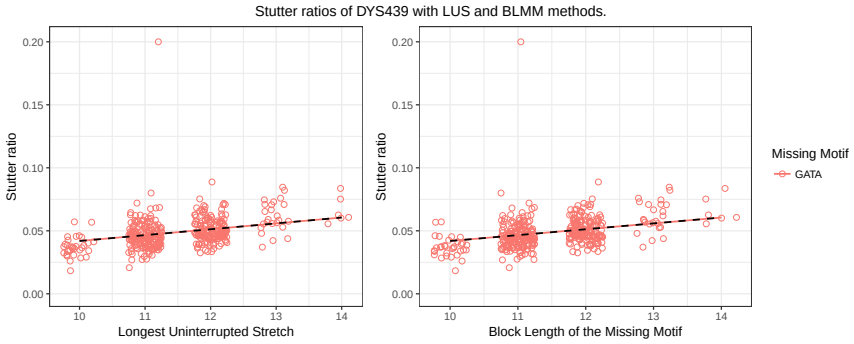

Fig. B.48: The stutter ratio plotted against the LUS and BLMM of DYS439. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

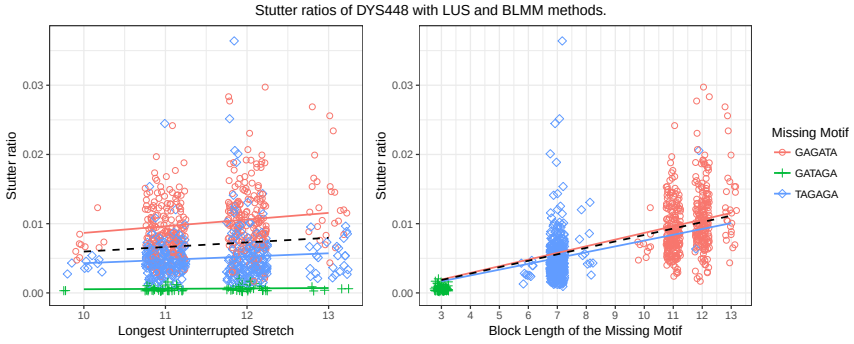

Fig. B.49: The stutter ratio plotted against the LUS and BLMM of DYS448. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

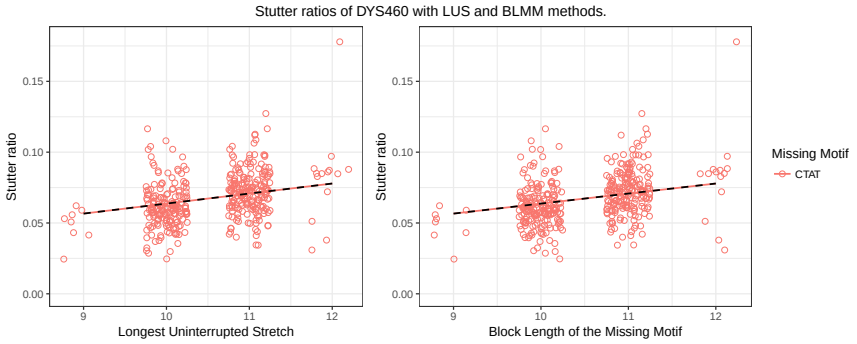

Fig. B.50: The stutter ratio plotted against the LUS and BLMM of DYS460. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

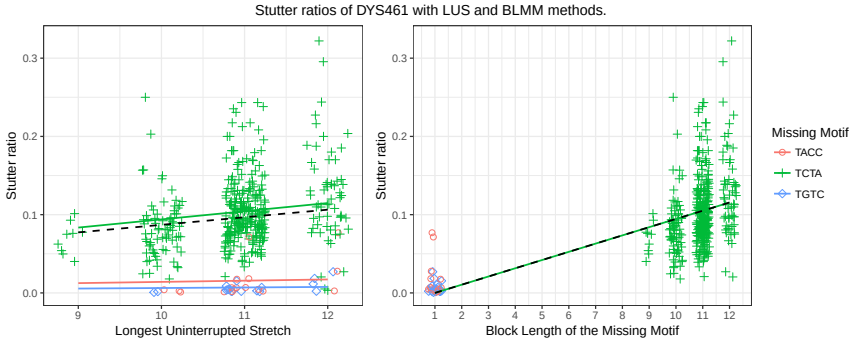

Fig. B.51: The stutter ratio plotted against the LUS and BLMM of DYS461. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

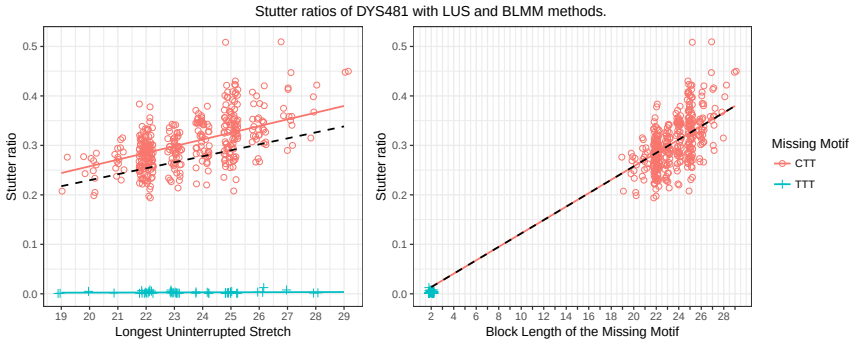

Fig. B.52: The stutter ratio plotted against the LUS and BLMM of DYS481. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

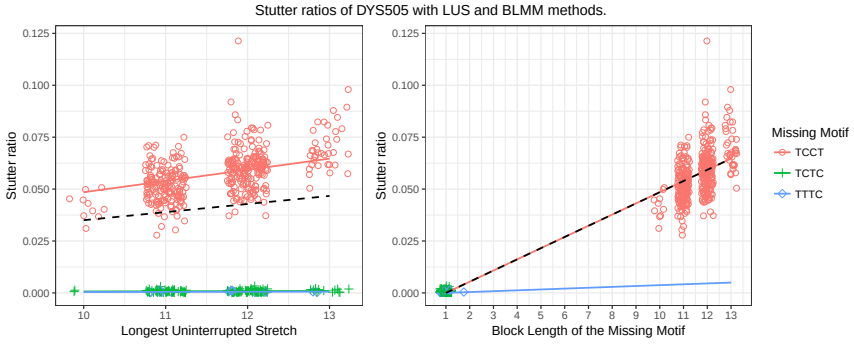

Fig. B.53: The stutter ratio plotted against the LUS and BLMM of DYS505. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

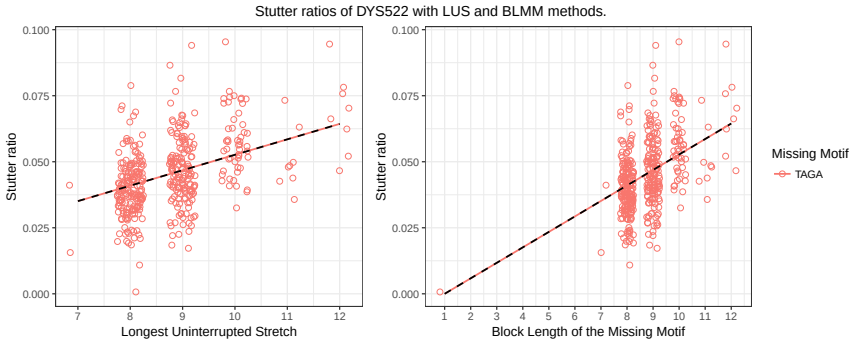

Fig. B.54: The stutter ratio plotted against the LUS and BLMM of DYS522. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

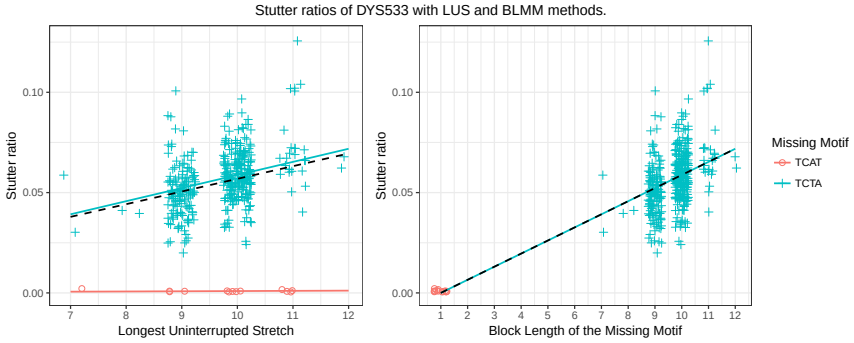

Fig. B.55: The stutter ratio plotted against the LUS and BLMM of DYS533. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

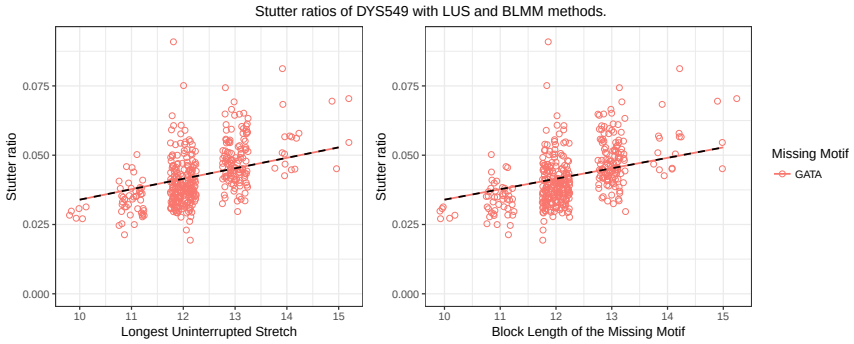

Fig. B.56: The stutter ratio plotted against the LUS and BLMM of DYS549. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

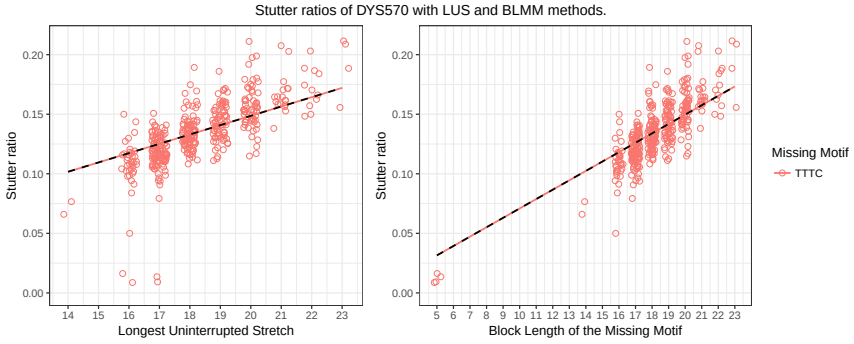

Fig. B.57: The stutter ratio plotted against the LUS and BLMM of DYS570. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

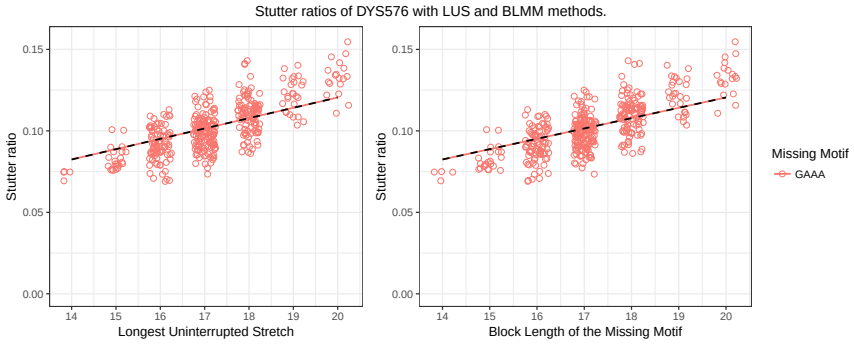

Fig. B.58: The stutter ratio plotted against the LUS and BLMM of DYS576. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

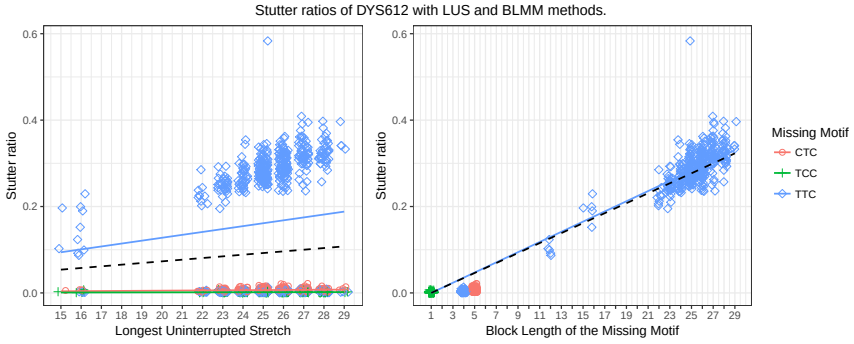

Fig. B.59: The stutter ratio plotted against the LUS and BLMM of DYS612. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

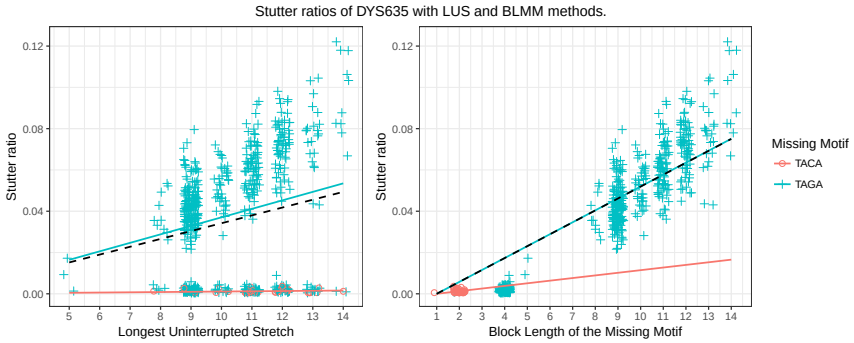

Fig. B.60: The stutter ratio plotted against the LUS and BLMM of DYS635. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

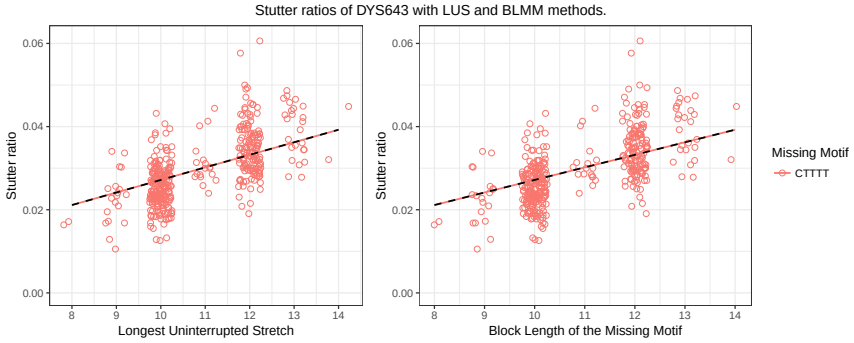

Fig. B.61: The stutter ratio plotted against the LUS and BLMM of DYS643. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

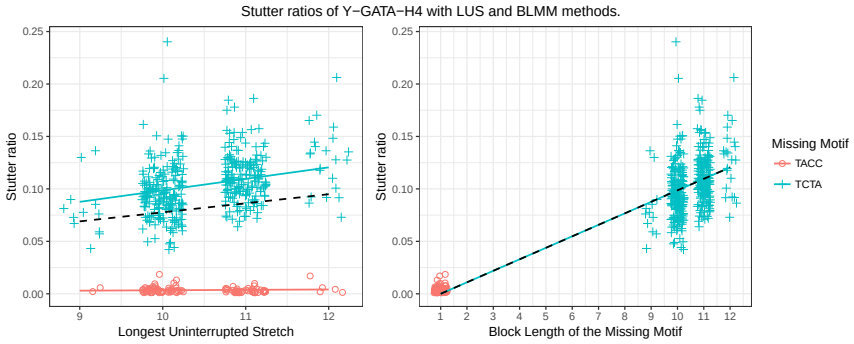

Fig. B.62: The stutter ratio plotted against the LUS and BLMM of Y-GATA-H4. The plots are coloured according to the missing motifs when the stutter sequence were compared to those of the parental alleles. The solid lines correspond to linear models with intercept through zero and slope dependent on both the markers and the missing motifs. The black dashed lines correspond to the linear model with intercepts through zero and marker dependent slopes. Note that a jitter has been added to visualise the density of the points.

References

[1] A. Edwards, A. Civitello, H. Hammond, and C. Caskey, "DNA typing and genetic mapping with trimeric and tetrameric tandem repeats," *American Journal of Human Genetics*, vol. 49, pp. 746–756, 1991.

[2] G. Levinson and G. Gutman, "Slipped-strand mispairing: A major mechanism for dna sequence evolution," *Molecular Biology and Evolution*, vol. 4, no. 3, pp. 203–221, 1987.

[3] X. Hauge and M. Litt, "A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR," *Human Molecular Genetics*, vol. 2, pp. 411–415, 1993.

[4] P. Walsh, N. Fildes, and R. Reynolds, "Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA," *Nucleic Acids Research*, vol. 24, pp. 2807 – 2812, 1996.

[5] M. Meldgaard and N. Morling, "Detection and quantitative characterization of artificial extra peaks following polymerase chain reaction amplification of 14 short tandem repeat systems used in forensic investigations," *Electrophoresis*, vol. 18, no. 11, pp. 1928–1935, 1997.

[6] K. Lazaruk, J. Wallin, C. Holt, T. Nguyen, and P. Walsh, "Sequence variation in humans and other primates at six short tandem repeat loci used in forensic identity testing," *Forensic Science International*, vol. 119, no. 1, pp. 1 – 10, 2001.

[7] M. Klintschar and P. Wiegand, "Polymerase slippage in relation to the uniformity of tetrameric repeat stretches," *Forensic Science International*, vol. 135, pp. 163 – 166, 2003.

[8] C. Brookes, J. Bright, S. Harbison, and J. Buckleton, "Characterising stutter in forensic STR multiplexes," *Forensic Science International: Genetics*, vol. 6, pp. 58 – 63, 2012.

[9] J.-A. Bright, D. Taylor, J. Curran, and J. Buckleton, "Developing allelic and stutter peak height models for a continuous method of DNA interpretation," *Forensic Science International: Genetics*, vol. 7, pp. 296 – 304, 2013.

[10] S. Vilsen, T. Tvedebrink, H. Mogensen, and N. Morling, "Statistical Modelling of Ion PGM HID STR 10-plex MPS Data," *Forensic Science International: Genetics*, 2017.

[11] W. Parson, D. Ballard, B. Budowle, J. M. Butler, K. B. Gettings, P. Gill, L. Gusmão, D. R. Hares, J. A. Irwin, J. L. King, P. D. Knijff, N. Morling, M. Prinz, P. M. Schneider, C. V. Neste, S. Willuweit, and C. Phillips, "Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements," *Forensic Science International: Genetics*, vol. 22, pp. 54–63, 2016.

[12] C. Hussing, C. Huber, R. Bytyci, N. Morling, and C. Børsting, "The Danish STR sequence database: Duplicate typing of 363 Danes with the ForenSeq^{TM} DNA Signature Prep Kit," *Int. J. Legal Med.*, 2018.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2009.

[14] A. Woerner, J. King, and B. Budoqle, "Flanking Variation Influences Rates of Stutter in Simple Repeats," *Genes*, vol. 8, 2017.

[15] D. Taylor, J. Bright, C. McGoven, C. Hefford, T. Kalafut, and J. Buckleton, "Validating multiplexes for use in conjunction with modern interpretation strategies," *Forensic Science International: Genetics*, vol. 20, pp. 6–19, 2016.

Paper C

Modelling allelic drop-outs in STR sequencing data generated by MPS

Søren B. Vilsen, Torben Tvedebrink, Poul Svante Eriksen, Claus Børsting, Christian Hussing, and Niels Morling

The paper has been published in the
Forensic Science International: Genetics, vol. 37, pp. 6–12, 2018.

The layout has been revised.

Abstract

We used a Poisson-gamma model to analyse the allele coverage of autosomal short tandem repeat (STR) systems obtained by massively parallel sequencing (MPS). The Poisson-gamma coverage model was created using the peak height models from capillary electrophoresis (CE) based detection of PCR products as a starting point. The CE models were modified to account for the differences between CE and MPS signals by accounting for the large marker imbalances seen for MPS data and by using the Poisson-gamma distribution instead of the normal, log-normal, or gamma distributions that were applied for CE data. We took two approaches to estimate the marker imbalance parameters by (1) using a work-flow data base, and (2) using the results of replicate investigations of the samples.

The Poisson-gamma model was used to estimate the rate of drop-outs of (1) single contributor dilution series experiments and (2) the minor contributor in two-person mixture samples. We examined the predictive capabilities of the model by comparing the observed and expected Brier scores of each sample. We derived the expected Brier scores and their variances to create asymptotic confidence intervals of the Brier scores. We found that the Poisson-gamma model performed well when using the work-flow data base, but that the replicate approach is not necessarily a viable option.

1 Introduction

DNA recovered at a crime scene is often found in low quantity and may be highly degraded. This affects the efficiency of the PCR and may lead to partial profiles with frequent locus and allele drop-outs, and in some situations also allele drop-ins, due to stochastic effects. [1–13]. We will only consider drop-outs. Typically, forensic genetics laboratories define a minimum threshold for acceptance of an allele. If this threshold is not reached, the allele is not detected and the DNA profile will suffer from allele drop-out. Accurate estimation of the probability of allelic drop-out is important since allele drop-outs will lead to mismatches between the DNA profile of the trace sample and the reference sample from a victim, a potential suspect, or an unrelated third party. The methods developed for assessing allelic drop-out of amplified PCR products detected by capillary electrophoresis (CE) cannot be used directly for massively parallel sequencing (MPS) assays. Thus, the methods need to be modified for use in the MPS setting.

Estimation of the probability of allelic drop-out in CE has generally taken one of two forms: (1) direct marker dependent estimation by logistic regression [14, 15], or (2) modelling the peak height signal and calculating the probability of being smaller than the threshold under the model [16–20].

The logistic regression model relied on H, the average peak height [14, 15]. In a previous paper [21], we concluded that the average allele coverage (equivalent to the average peak height in CE) was not an apt covariate of the probability of drop-out because the libraries were normalised prior to sequencing. While this was the case for the Ion PGM 10-plex data, it was not necessarily the case for the data generated by the Illumina® ForenSeq™ Signature Prep Kit. In our previous paper [21], we suggested the standard deviation of the heterozygote balance as a possible replacement. This hypothesis was tested during the preliminary data analysis. However, the performance was deemed too poor and, thus, not included.

Therefore, in order to estimate the probability of allelic drop-out, we will model the allele coverage. Many of the lessons learned by modelling peak heights in CE should translate when modelling the coverage,because the data are based on analysis of PCR products. As

peak heights are inherently continuous (though often round to nearest integer), they have been modelled using normal, log-normal, and gamma distributions [16–20]. Even though the choice of distribution differs, these models exhibit similar mean and variance structures. We have chosen to focus on the 'gamma model' [18, 22, 23], which assumes that for sample s and the marker m, the peak height of allele a, H_{sma}, follows a gamma distribution with mean and variance given as:

$$\mathbb{E}\left[H_{sma}\right] = \eta_s \rho_s \sum_{c=1}^{C} \{(1-\xi_s) g_{smac} + \xi_s g_{sm(a+1)c}\} \varphi_c, \quad \text{and} \tag{C.1}$$

$$\mathbb{V}\mathrm{ar}\left[H_{sma}\right] = \eta_s \mathbb{E}\left[H_{sma}\right], \tag{C.2}$$

respectively, where c is a contributor to the mixture, C is the total number of contributors to the mixture, ρ_s is proportional to the amount of input DNA, φ_c is the mixture proportion of individual c, ξ_s is the average stutter proportion of the sample, and g_{smac} is the number of alleles a that contributor c has. Thus, g_{smac} is 0, 1, or 2 if the contributor c does not have any allele a, is heterozygous, or homozygous for allele a, respectively.

If the coverage was large enough (for all contributors), the coverage could be modelled using the CE methods, as the continuous models would provide a satisfactory approximation of the discrete coverage. However, if the DNA of the sample is of low quantity, if the sample is degraded, or if the DNA mixture proportions is skewed (e.g. with DNA from one contributor in low quantity compared to that of the major contributor to the mixture), a continuous model would lead to a poor approximation.

Using the mean structure of the gamma model as a starting point, we modelled the allele coverage by a Poisson-gamma distribution (also known as a negative binomial distribution parameterised by its mean). We made modifications to the mean structure because: (1) The MPS data suffers from large marker imbalances, which was accounted for by multiplying the expected coverage by a marker dependent scaling factor, and (2) parental alleles can produce multiple stutters, and a stutter can have more than one possible parental allele [24].

If thresholds have not been applied, the coverage model presented in this manuscript quite naturally accounted for drop-out by its definition. However, as thresholds limit the amount of information, we

took two slightly different approaches to estimate the marker dependent parameters of the model, to ensure we had enough information to estimate all the parameters. The first approach relied on a database of reference samples (a workflow database), while the second used multiple replicates of the sample. Given the estimated parameters, we found the probability of allelic drop-out by calculating the probability of the coverage being smaller than the threshold.

The aim of this manuscript is to present an MPS DNA mixture coverage model, called the Poisson-gamma coverage model, and investigate its behaviour when used to analyse DNA samples. The DNA samples analysed in this manuscript were described in Hussing et al. [25]. They also described the marker imbalances, marker drop-out rate, and other characteristics of the samples.

In order to measure the performance of the presented model, we examined the how well the coverage model predicted the probability of drop-out. We compared the observed drop-out with the probability of drop-out predicted under the model using the Brier-score. The observed Brier-score was compared to the expected Brier-score in order to investigate whether the coverage model behaved as expected. We believe that the model and thoughts presented in the manuscript may serve as a foundation for the work going forward in modelling STR DNA mixture samples.

2 Materials and methods

2.1 Experimental data

DNA libraries were build using the ForenSeq™ DNA Signature Prep Kit (Illumina®) Primer Mix A and B. Primer Mix A amplifies markers for human identification (HID), while Primer Mix B amplifies the same HID markers plus ancestry informative markers (AIMs) and markers associated with eye and hair colour. DNA sequencing was performed with the MiSeq FGx (Illumina®) as previously described in [25, 26].

DNA was extracted from blood samples and buccal swabs collected on FTA cards from 363 individuals. The samples were amplified and sequenced in duplicate.

Dilution series of DNA were created from four contributors. The DNA was amplified and sequenced in triplicate using Primer Mix A.

The amounts of DNA in each series were 1 ng, 500 pg, 250 pg, 125 pg, 62.5 pg, 31.25 pg, 15.63 pg, and 7.86 pg. A consensus DNA profile from each individual was generated based on all experiments with 1 ng input.

Fifteen two person DNA mixture samples (in proportions: 1:1,000, 1:100, 1:50, 1:25, 1:12, 1:6, 1:3, 1:1, 3:1, 6:1, 12:1, 25:1, 50:1, 100:1, and 1,000:1) were made with a male and a female contributor. The total amount of DNA was 1 ng in all cases. The DNA mixtures were amplified and sequenced in duplicate using Primer Mix B. The profiles of the two contributors were known.

Sequences were identified by their flanking regions using STRait Razor v3.0 [27]. They were subsequently trimmed to include just the STR region and the coverage of every unique sequence found. Thus, if two sequences had the same length, but their sequences were different (sometimes referred to as isoalleles), they were treated as different sequences throughout the manuscript.

2.2 Analytic thresholds

We opted for simple marker dependent thresholds, taking a percentage, p, of the maximum coverage of the marker [21]. That is, the threshold of marker m from sample s is:

$$t_{sm} = p \cdot \max_a \{y_{sma}\}, \tag{C.3}$$

where y_{sma} is the coverage of sample s, marker m, and allele a. Thus, given a marker where we have observed three strings a, b, and c, with coverage 100, 90, and 4, respectively, i.e. the maximum coverage on the marker is 100. Assuming that $p = 0.05$, then the threshold $t = 5$, implying that the observation c was classified as noise.

We chose this method for establishing the thresholds because it was the simplest method for defining marker dependent thresholds. This method does not change the modelling approach, the determination of the probability of drop-out (other than slight changes to the probabilities, themselves), or the conclusions.

2.3 The coverage model

The coverage of an allele is not continuous. The coverage of a string is a synonym for the count of the string. Thus, it would be natural

to model the coverage as count data, i.e. discrete. The most common, and simplest, choice of model for count data is the Poisson distribution that assumes equal mean and variance, which is unrealistic for MPS STR data. We modelled the coverage by using a Poisson-Gamma distribution (see Appendix A), which includes a dispersion parameter to account for the possible deviations from the assumption of equal mean and variance.

We introduce the mean structure of the coverage model in three stages, by (1) introducing the model corresponding to the situation, where there is only one contributor and no artifact, (2) accounting for multiple contributors, and (3) accounting for stutters.

2.3.1 The mean structure of the coverage model

Under the assumption that sample s was a single contributor and without any artefact, i.e. the sample did not show drop-outs, drop-ins, or stutters. The mean structure of the gamma model seen in Eq. (C.1) is reduced to $\eta_s \rho_s g_{sma}$. We wanted the Poisson-gamma model to behave in a similar fashion. However, the MPS process suffers from large marker imbalances. Therefore, we defined the mean of the single-contributor Poisson-gamma model with no artefacts, $\tilde{\mu}_{sma}$, as

$$\tilde{\mu}_{sma} = \nu_s \beta_m g_{sma}, \tag{C.4}$$

where β_m is the marker imbalance parameter, the average of the marker imbalance parameters is equal to one in order to avoid overspecification of the model (i.e. $(\sum_m \beta_m)/M = 1$), and ν_s could be interpreted as the average coverage of the heterozygotes of sample s.

However, the genotype of a crime scene stain is potentially a mixture of genotypes from multiple contributors. Thus, we need to extend the mean structure in Eq. (C.4) to handle multiple contributors. We denoted the genotype and relative contribution to the mixture of contributor, c, as g_{smac} and φ_c, respectively, with $0 < \varphi_c < 1$ and $\varphi_+ = \sum_c \varphi_c = 1$. Thus, for allele a, we get a sum of the contributor's genotypes weighted by their relative contributions to the mixture:

$$\sum_{c=1}^{C} g_{smac} \varphi_c, \tag{C.5}$$

assuming a total of C contributors to the mixture.

Stutter products are a common artefact of the PCR amplification of STRs and, thus, the rate with which stutters are created needed to be incorporated into the model. Two measures of the stutter rate are commonly used: stutter ratio and stutter proportion. We used the stutter ratio for modelling purposes. Because of the added resolution of the MPS process, a stutter sequence can have received coverage from multiple parental alleles. Therefore, we needed a predictor of the stutter ratio, which could handle the added resolution. The block length of the missing motif (BLMM) is such a predictor [24]. The BLMM was created as an extension of the longest uninterrupted stretch (LUS). However, instead of using the longest stretch, we find the stretch from which the parental allele had lost a motif when compared to the stutter. The length of the found stretch is the BLMM.

It has been shown [24] that a suitable model for the relationship between the stutter ratio and the BLMM is linear through the point $(1, 0)$, i.e.

$$\mathrm{SR} = \alpha_m \cdot (\mathrm{BLMM} - 1), \tag{C.6}$$

where α_m are marker dependent slope parameters of the linear model. The slope parameters were estimated on a database of samples sequenced using the same technology, panel, and settings (i.e. the same number of PCR cycles, preparation, etc.), called a workflow database, as the analysed samples.

As a stutter can have multiple parental alleles (imposed by the higher resolution of the MPS process), the very simple term accounting for the stuttering in Eq. (C.1), $\xi_s g_{sm(a+1)c}$, needs to be extended to a sum over the set of parental alleles of sequence a, $\mathcal{P}(a)$, i.e.

$$\sum_{A \in \mathcal{P}(a)} \xi_{mA}\, g_{smAc},$$

where A is a parental sequence of the sequence a and ξ_{mA} is the predicted stutter ratio of the parental sequence A.

It follows that we can write contributor c's contribution to allele a as

$$\left(g_{smac} + \sum_{A \in \mathcal{P}(a)} \xi_{mA}\, g_{smAc} \right) \varphi_c. \tag{C.7}$$

2.4 Estimation of parameters

Assuming that the sample contains a single contributor, no stutter reads, no drop-in, and no drop-out, the expected coverage of the model takes the form described in Eq. (C.4). This model requires $M+1$ parameters ($M-1$ accounting for marker imbalances, one for the parameter corresponding to the average coverage of heterozygote alleles, and one for the dispersion parameter). A sample with a single contributor would have between M and $2M$ alleles. That is, if the sample did not have any drop-outs, we would in most (if not all) practical applications have enough observations for the parameters to be estimated by maximising the likelihood (MLE). However, if the sample suffers from allele (and/or marker) drop-out, this would no longer be guaranteed.

Preliminary analyses (not included in this manuscript) showed that the marker imbalances were consistent across samples, batches, and preparation kits. In order to reduce the number of parameters, which needed to be estimated, we studied two slightly different approaches of utilising this consistency:

(1) **The reference approach:** A workflow database is used to estimate the marker dependent parameters. The marker dependent parameters in the reference approach were fitted to the model with mean structure in Eq. (C.4) using a database samples. These parameters were used in subsequent analyses, which reduced the number of parameters from $M+1$ to two per sample (as the $M-1$ marker imbalances are considered known with this approach).

(2) **The replicate approach:** The R replicates of the sample were collected, and the $M-1$ marker dependent parameters were assumed to be equal among all R replicates. This reduced the number of parameters from $R(M+1)$ to $2R+(M-1)$ for the R replicates. Note $R(M+1)$ is equal to $2R+(M-1)$ only if $R=1$ or $M=1$.

The reference approach utilises the marker information from the technology as a whole, while the replicate approach only uses the information of the replicates. The reference approach is more stable, while the replicate approach is more flexible. The replicate approach can be used for any sample using any MPS technology as it would only depend on the replicates, while the reference approach would always need an entire database of sample data sequenced with the same

technology, panel, and settings (e.g. the same number of PCR cycles, preparation, and so on) as the sample in question, referred to as a workflow database. The workflow database used in this manuscript consisted of the 363 Danes sequenced in duplicate presented in Hussing et al. [26].

2.5 Assessment of predictive capabilities

The predictive capabilities of the models were assessed by the Brier-score [28]. Given the observed vector of drop-out, $\boldsymbol{d} = (d_{11}, ..., d_{M\mathcal{A}_M})^T$ with $d_{ma} = \mathbb{I}[y_{sma} \leq t_{sm}]$, and the predicted probability of drop-out, $\hat{\boldsymbol{p}} = (\hat{p}_{11}, ..., \hat{p}_{M\mathcal{A}_M})^T$ with $\hat{p}_{ma} = \mathbb{P}(y_{sma} \leq t_{sm} | \mu_{sma}, \eta_s)$ of a sample, s, the Brier-score of sample s is defined as:

$$\mathcal{B}_s(\boldsymbol{d}, \hat{\boldsymbol{p}}) = \frac{1}{\mathcal{A}_+} \sum_m \sum_a (d_{ma} - \hat{p}_{ma})^2, \tag{C.8}$$

where $\mathcal{A}_+ = \sum_m \mathcal{A}_m$, and $\mathcal{A}_m$ is the number of alleles on marker m. Note, that Eq. (C.8) is a special case of the Brier-score, also called *'the half Brier score'* or *'the mean square error'*. In order to assess the performance of the methods, we plotted the observed Brier score against the expected Brier score with point-wise confidence intervals for each sample. The derivation of the expected Brier score, the variance of the Brier score, and the confidence intervals are shown in Appendix B.

Lastly, we should note that the Brier-score has a tendency to become more insensitive as the events (probability of drop-out) become more extreme [29]. This resulted in an increase in the variance of the Brier-score and, therefore, larger confidence intervals, as the probability of drop-out tended to zero (or one).

3 Results

We analysed the data (1) without analytic thresholds, and (2) with analytic thresholds using the method described in Section 2.2. The first approach is the best case scenario under which the model has to perform well to be a viable option. The second approach tests the model in conditions that are closer to real case work scenarios. In the case of the former, the coverage of an allele is set to zero if not observed (i.e. if it dropped out).

3.1 Dilution series experiment

For each of the four individuals, who contributed to the dilution samples, we created a consensus profile. This consensus profile was used as the known genotype in the dilution series and replicates thereof.

The first column in Figure C.1 shows the observed Brier score against the expected Brier score for each sample in the dilution series experiments using the reference approach to estimate the marker dependent parameters. The shaded 95% confidence areas in the figure were created on a sample-by-sample basis connecting the lower and upper limits of the confidence intervals of the samples. Without thresholds, the observed Brier score of one sample was found outside the confidence interval, i.e. approximately 99% of the results were within their 95% confidence limits. However, if thresholds were applied, the number of samples who had observed Brier scores outside the confidence limits increased to six. As noted above, this difference was expected because of the loss of information.

The second column of Figure C.1 shows the observed Brier score against the expected Brier score for the replicate approach. If thresholds were not used, a single sample had an observed Brier score outside its confidence limits, i.e. approximately 99% of the samples were inside the their confidence limits. Compared to 12 samples outside their confidence limits when thresholds were used (approximately 88% inside their confidence limits). The 12 samples who had observed Brier scores outside their confidence limits were samples with small amounts of input DNA (four with 7.81 pg, one with 31.25 pg, five with 62.50 pg, and two with 125 pg). This was in part a consequence of the added variability in allele coverage. However, it was mainly due to the loss of information introduced by applying thresholds.

3.2 Mixture samples

For DNA mixture samples, the drop-out rate of interest is that of the minor contributor to the mixture. Thus, after parameter estimation, we limited the calculation of the observed and expected Brier scores to the alleles of the minor contributor that were not shared with the major contributor. Note that isoalleles (alleles of the same length, but with different genetic sequences) were treated as being different, i.e. not shared in case mixtures.

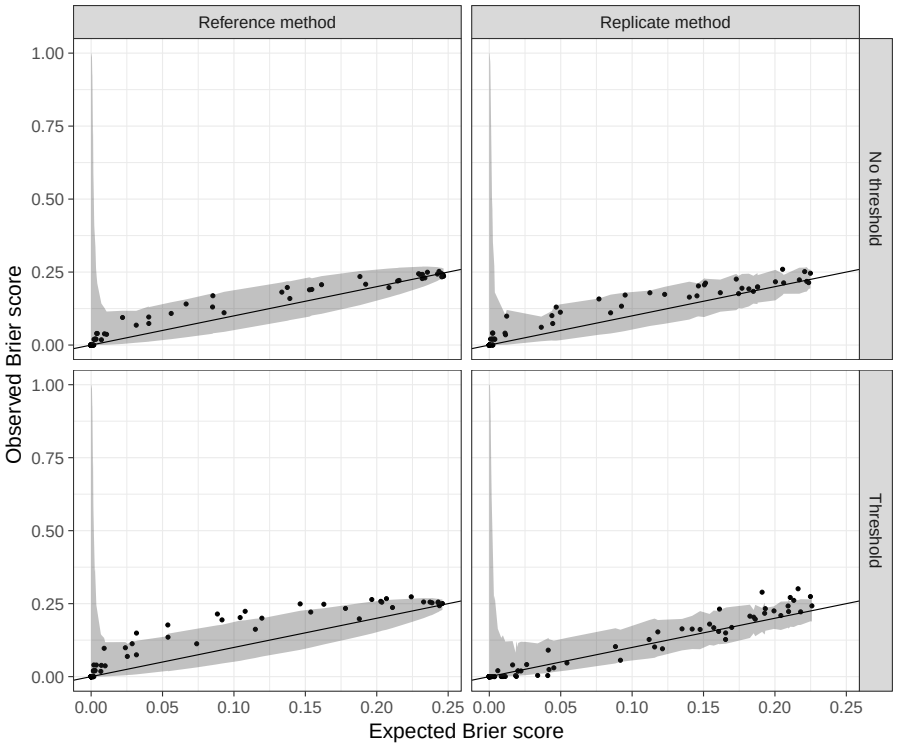

Fig. C.1: The observed and expected Brier scores for all 96 samples in the dilution series. The solid line is the one-to-one relationship between the observed and expected Brier scores. The shaded areas were created by connecting the confidence intervals of each sample. The marker dependent parameters of the Poisson-gamma model were estimated using the reference and replicate approaches shown in the first and second column, respectively.

The observed Brier score against the expected Brier score for the mixture samples can be seen in the first and second column of Figure C.2 using the reference and replicate approaches, respectively. In both cases, a single sample was found to exceed its confidence limits when thresholds were not applied versus four samples when using thresholds to censor the coverage (though it was not the same four samples in the replicate as in the replicate approach). Furthermore, we saw that even though only a single sample falls outside the confidence bounds when not using thresholds, the expected Brier score was much smaller than when using thresholds, in both the reference and replicate approaches. The difference in the observed and expected

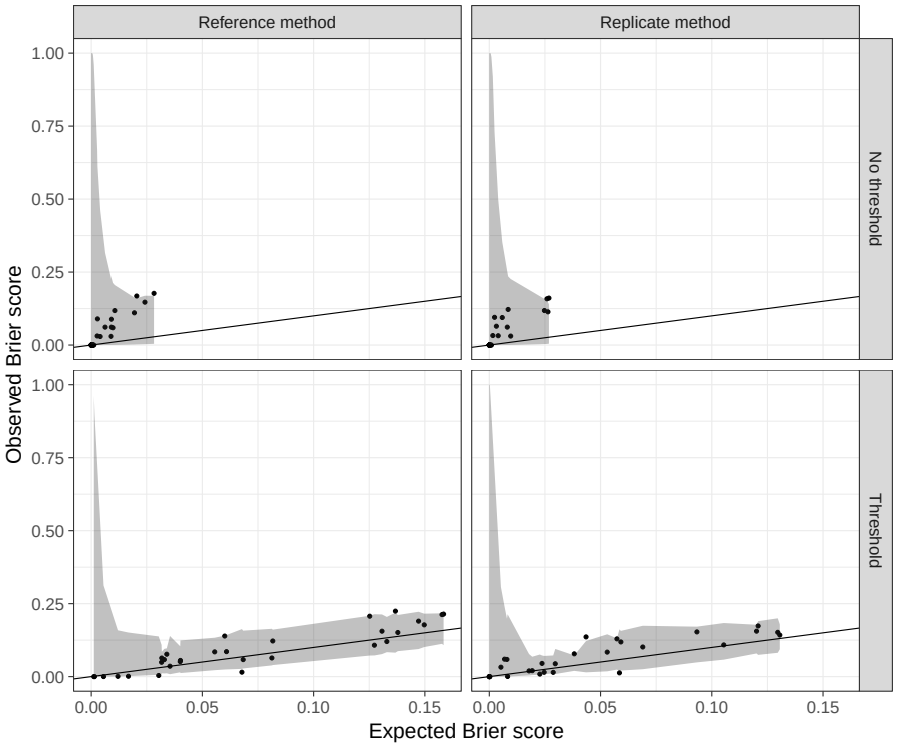

Fig. C.2: The observed and expected Brier scores for all 30 mixture samples. The observed and expected Brier scores were restricted to alleles of the minor contributor not shared with the major contributor. The solid line is the one-to-one relationship between the observed and expected Brier scores. The shaded areas were created by connecting the confidence intervals of each sample. The marker dependent parameters of the Poisson-gamma model were estimated using the reference and replicate approaches shown in the first and second column, respectively.

Brier score, is most likely a consequence of the number of drop-outs being much smaller, when not using thresholds. This will decrease the average probability of drop-out and, thereby, the expected Brier score.

Note that even though the samples were within the confidence limits, when thresholds were not applied, their observed Brier scores appeared slightly biased when compared to the expected Brier scores, in both the dilution and mixture experiments.

4 Conclusion

The primary consequence of treating all isoalleles (strings of the same length, but with genetically different sequences) as different alleles was an increased number of allele drop-outs. This is a simple consequence of observing a larger number of heterozygotes due to the increased resolution of the MPS process. As heterozygote alleles have lower coverages, than homozygote alleles it follows that they will have larger probabilities of falling below a given threshold for samples with low quantities of template DNA (or for contributors in DNA mixtures contributing a small amount of DNA to the sample).

Using a workflow database, the reference approach created more robust estimates of the marker imbalance parameters than the replicate approach if we applied thresholds. However, if thresholds were not used, we saw nearly no difference between the reference and replicate approaches. This is a simple consequence of using thresholds as these more frequently create drop-out, because they censor information. Thus, from a modelling perspective: if thresholds were applied to the coverage, we would prefer to use a work-flow database when estimating the marker imbalance parameters.

The main advantage of the replicate approach is that it would only need the replicates of the sample of interest. Thus, would require much less preparation, or would it? The answer is: it depends. It depends on whether the laboratory sequencing the samples has made internal calibration and validation of the machines (PCR, sequencing, etc.) and kits used to perform the sequencing. If they have, then the samples sequenced during validation can be used as a workflow database. The samples used to create the workflow database would not need to be obtained from dilution series experiments, though these could be included if needed as long as the DNA profile of the sample donor is known. The workflow database can be extended over time, as any sample even from unknown donors can be used. Thus, after creating e.g. reference samples or samples used for parentage testing, the resulting sequenced samples can be added to the workflow database. Furthermore, the workflow database would also be needed to estimate the slope parameters of the (BLMM) stutter model [24]. Thus, having a workflow database would still be implicitly necessary even if we used the replicate approach.

In creating the Poisson-gamma coverage model, we matched the mean as close to current CE methods as possible. However, the variance of the Poisson-gamma model is:

$$\mathbb{V}\mathrm{ar}\left[Y_{sma}\right] = \mu_{sma}(1 + \eta_s^{-1}\mu_{sma}), \tag{C.9}$$

as seen in Eq. (C.12). That is, the standard deviation is approximately proportional to the mean for fixed η_s:

$$\mathrm{sd}(Y_{sma}) = (\mu_{sma} + \eta_s^{-1}\mu_{sma}^2)^{1/2} \approx (\eta_s^{-1}\mu_{sma}^2)^{1/2} = \eta_s^{-1/2}\mu_{sma},$$

for large values of μ_{sma}.

The most common CE peak height models define the mean to be proportional to the variance (in case of the gamma model see Eq. (C.2), but is also true for the log-normal model used by STRmix [17]). We can create an equivalent Poisson-gamma model by assuming

$$\mathbb{V}\mathrm{ar}\left[Y_{sma}\right] = \mu_{sma}(1 + \theta_s), \tag{C.10}$$

where θ_s is the new overdispersion parameter. This parameterisation is usually called the Poisson-gamma model of order 1, or PG1 model, because the variance is proportional to the mean to the power of 1. Using similar reasoning, the Possion-gamma coverage model in Eq. (C.9) is called the PG2 model, as the variance is approximately proportional to the mean squared. We compared the performances of the PG1 and PG2 models to each other, as described in Section 2.5. Figure C.3 shows the observed Brier scores against the expected Brier scores for the PG1 and PG2 models. Focusing on the blue coloured points and shading, the figures shows that the PG1 model is inferior to the PG2 model whether thresholds were applied or not.

During the preliminary analyses, early results showed that the coverage of homozygotes might not necessarily be twice that of heterozygotes, and that this effect might be marker dependent. Therefore, we also tried to estimate a marker dependent homozygote scaling factor. The Brier-scores of the estimated models are shown in Figure C.3. The red and blue coloured points (and shadings) correspond to the case where the homozygote scaling factors were estimated using a reference approach, and the case where the homozygote scaling was set to two for all markers, respectively. It is evident that for the PG2 model, there is no difference between the predictive capabilities whether the

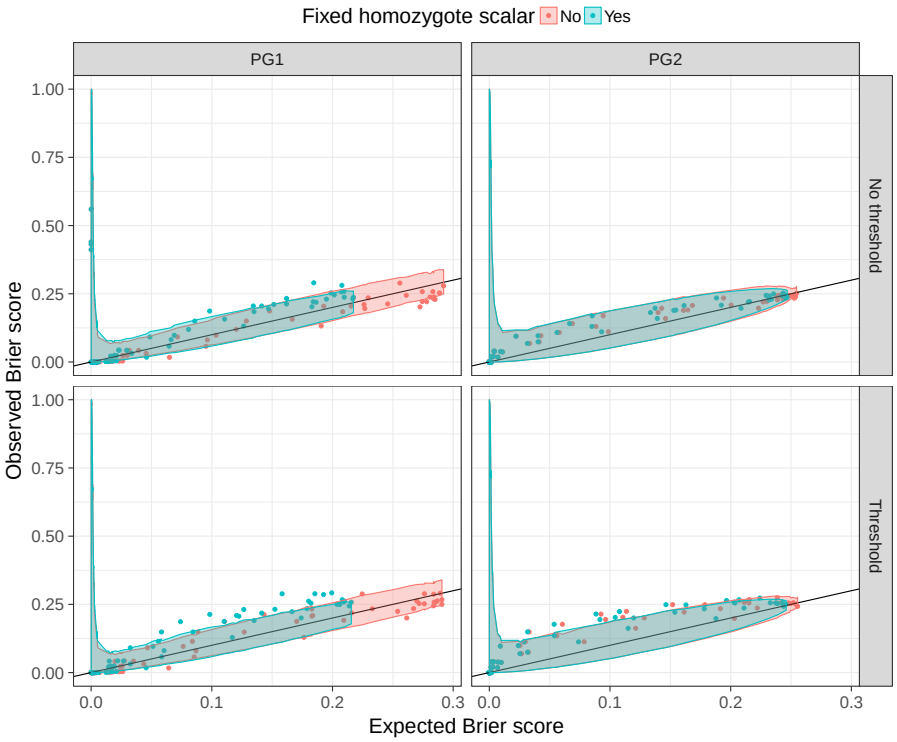

Fig. C.3: The observed and expected Brier scores of 96 samples in a dilution series. The solid line is the one-to-one relationship between the observed and expected Brier scores. The shaded areas were created by connecting the confidence intervals of each sample. The marker dependent parameters of the Poisson-gamma model were estimated using the reference approach. The samples were fitted using the PG1 and PG2 models shown on the left and right column, respectively. The bottom and top rows shows results with and without thresholds, respectively. The red and blue dots shows the Brier scores when estimating the homozygote scalars and fixing them at 2, respectively.

homozygote scaling factor was estimated or set to two. We will note that, if the homozygote scaling was different from two, it would imply that the coverage is not necessarily additive. This implication would not just effect homozygotes. When analysing a two person mixture, if both contributors have a common allele, the coverage of said allele would not be the sum of the individual coverages provided by each of the contributors. That is, if this was true, then it would have disastrous consequences for the usage of the MPS technology. However, as

we found no difference in the predictive capabilities and we did not find any biochemical reason for this behaviour, we recommend using the PG2 model with a fixed homozygote scaling factor of two.

Appendices

A The Poisson-gamma distribution

Assuming that the coverage of sample s, marker m, allele a, denoted Y_{sma}, follows a Poisson-Gamma distribution with mean μ_{sma} and dispersion η_s, then the probability of observing y_{sma} is given as:

$$\mathbb{P}\left(Y_{sma} = y_{sma} | \mu_{sma}, \eta_s\right) = \frac{\Gamma(y_{sma} + \eta_s)}{\Gamma(y_{sma} + 1)\Gamma(\eta_{sma})} \frac{\mu_{sma}^{y_{sma}} \eta_s^{\eta_s}}{(\mu_{sma} + \eta_s)^{y_{sma} + \eta_s}},$$

where $\Gamma(x)$ is the gamma function. The mean and variance of Y_{sma} are given as:

$$\mathbb{E}\left[Y_{sma}\right] = \mu_{sma} \quad \text{and} \tag{C.11}$$

$$\mathbb{V}\text{ar}\left[Y_{sma}\right] = \mu_{sma}(1 + \eta_s^{-1}\mu_{sma}). \tag{C.12}$$

B Expectation and variance of the Brier score

Assume d_n follows a Bernoulli distribution with parameter p_n for $n = 1, ..., N$, then the Brier score is given as:

$$\mathcal{B}(\boldsymbol{d}, \boldsymbol{p}) = \frac{1}{N} \sum_n (d_n - p_n)^2. \tag{C.13}$$

Furthermore, assuming that d_n is independent of d_m when $n \neq m$, then taking the expectation of Eq. (C.13) yields:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}(\boldsymbol{d}, \boldsymbol{p})\right] &= \frac{1}{N} \sum_n \mathbb{E}\left[(d_n - p_n)^2\right] \\
&= \frac{1}{N} \sum_n p_n(1 - p_n) \\
&= \bar{p}(1 - \bar{p}) - \frac{1}{N} \sum_n (p_n - \bar{p})^2.
\end{aligned}$$

The variance of the Brier score is:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\mathcal{B}(\boldsymbol{d},\boldsymbol{p})\right] &= \frac{1}{N^2}\sum_n \mathbb{V}\mathrm{ar}\left[(d_n-p_n)^2\right] \\ &= \frac{1}{N^2}\sum_n \mathbb{E}\left[(d_n-p_n)^4\right] - \mathbb{E}\left[(d_n-p_n)^2\right]^2 \\ &= \frac{1}{N^2}\sum_n p_n - 5p_n^2 + 8p_n^3 - 4p_n^4,\end{aligned}$$

where the first equality holds as we assumed d_n and d_m to be independent when $n \neq m$.

In order to ensure that the lower and upper confidence bounds stay between 0 and 1, we create the confidence interval on the logit-scale by using the delta approximation, and then transforming the created bounds back to the scale of the Brier score using the sigmoid transformation (the inverse of the logit transformation). Thus, the point-wise confidence interval is given by:

$$f^{-1}\left(f\left(\mathbb{E}\left[\mathcal{B}(\boldsymbol{d},\hat{\boldsymbol{p}})\right]\right) \pm \frac{z_{1-\alpha/2}}{f'\left(\mathbb{E}\left[\mathcal{B}(\boldsymbol{d},\hat{\boldsymbol{p}})\right]\right)}\left(\mathbb{V}\mathrm{ar}\left[\mathcal{B}(\boldsymbol{d},\hat{\boldsymbol{p}})\right]\right)^{1/2}\right), \quad \text{(C.14)}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution and $f(\cdot)$, $f'(\cdot)$, and $f^{-1}(\cdot)$ are the logit function, the derivative of the logit function, and the sigmoid function, respectively.

References

[1] P. Gill, J. Whitaker, C. Flaxman, N. Brown, and J. Buckleton, "An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA," *Forensic Sci. Int.*, vol. 112, pp. 17 – 40, 2000.

[2] P. Gill, J. Curran, and K. Elliot, "A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci," *Nucleic Acids Research*, vol. 33, pp. 632 – 643, 2005.

[3] L. Dixon, A. Dobbins, H. Pulker, J. Butler, P. Vallone, M. Coble, W. Parson, B. Berger, P. Grubweiser, H. Mogensen, N. Morling,

K. Nielsen, J. Sanchez, E. Petkovski, A. Carrecedo, P. Sanchez-Diz, E. Ramos-Luis, M. Brion, J. Irwin, R. Just, O. Loreille, T. Parsons, D. Syndercombe-Court, H. Schmitter, B. Stradmann-Bellinghausen, K. Bender, and P. Gill, "Analysis of artificially degraded DNA using STRs and SNPs - results of a collaborative European (EDNAP) exercise," *Forensic Sci. Int.*, vol. 164, pp. 33 – 44, 2006.

[4] P. Smith and J. Ballantyne, "Simplified low-copy-number DNA analysis by post-PCR purification," *J. Forensic Sci.*, vol. 52, pp. 820 – 829, 2007.

[5] L. Forster, J. Thomson, and S. Kutranov, "Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle "low copy number" (LCN) method for analysis of trace forensic DNA samples," *Forensic Sci. Int. Genet.*, vol. 2, pp. 318 – 328, 2008.

[6] A. Western, J. Nagel, C. Benschop, N. Weiler, B. de Jong, and T. Sijen, "Higher capillary electrophoresis injection settings as an efficient approach to increase the sensitivity of STR typing," *J. Forensic Sci.*, vol. 54, pp. 591 – 598, 2009.

[7] P. Gill and J. Buckleton, "A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number," *Forensic Sci. Int. Genet.*, vol. 4, pp. 221 – 227, 2010.

[8] S. Petricevic, J. Whitaker, J. Buckleton, S. Vintiner, J. Patel, P. Simon, H. Ferraby, W. Hermiz, and A. Russell, "Validation and development of interpretation guidelines for low copy number (LCN) DNA profiling in New Zealand using the AmpF'STR1 SGM plus^{TM} multiplex," *Forensic Sci. Int. Genet.*, vol. 4, pp. 305 – 310, 2009.

[9] C. Benschop, C. van der Beek, H. Meiland, A. van Gorp, A. Western, and T. Sijen, "Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results," *Forensic Sci. Int. Genet.*, vol. 5, pp. 316 – 328, 2011.

[10] S. Cowen, P. Debenham, A. Dixon, S. Kutranov, J. Thomson, and K. Way, "An investigation of the robustness of the consensus method of interpreting low-template DNA profiles," *Forensic Sci. Int. Genet.*, vol. 5, pp. 400 – 406, 2011.

[11] A. Western, L. Grol, J. Harteveld, A. Matai, P. D. Knijff, and T. Sijen, "Assessment of the stochastic threshold, back-, and forward stutter filters and low template techniques for NGM," *Forensic Sci. Int. Genet.*, vol. 6, pp. 708 – 715, 2012.

[12] C. P. R. Klein-Unseld, M. Klintschar, and P. Wiegand, "Comparison of different interpretation strategies for low template DNA mixtures," *Forensic Sci. Int. Genet.*, vol. 6, pp. 716 – 727, 2012.

[13] C. Børsting, H. Mogensen, and N. Morling, "Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples," *Forensic Sci. Int. Genet.*, vol. 7, pp. 345 – 352, 2013.

[14] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling, "Estimating the probability of allelic drop-out of STR alleles in forensic genetics," *Forensic Science International: Genetics*, vol. 3, pp. 222 – 226, 2009.

[15] T. Tvedebrink, P. S. Eriksen, M. Asplund, H. S. Mogensen, and N. Morling, "Allelic drop-out probabilities estimated by logistic regression - Further considerations and practical implementation," *Forensic Science International: Genetics*, vol. 6, pp. 263 – 267, 2012.

[16] T. Tvedebrink, M. Asplund, P. S. Eriksen, H. S. Mogensen, and N. Morling, "Estimating drop-out probabilities of STR alleles accounting for stutters, detection threshold truncation and degradation," *Forensic Science International: Genetics Supplement Series*, vol. 4, pp. e51 – e52, 2013.

[17] D. Taylor, J.-A. Bright, and J. Buckleton, "The interpretation of single source and mixed DNA profile," *Forensic Science International: Genetics*, vol. 7, pp. 516 – 528, 2013.

[18] R. Cowell, T. Graversen, S. Lauritzen, and J. Mortera, "Analysis of Forensic DNA Mixtures with Artefacts," *Royal Statistical Society. Journal Series C: Applied Statistics*, vol. 64, pp. 1 – 32, 2015.

[19] Ø. Bleka, G. Storvik, and P. Gill, "EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts," *Forensic Science International: Genetics*, vol. 21, pp. 35 – 44, 2016.

[20] C. Steele, M. Greenhalgh, and D. Balding, "Evaluation of low-template DNA profiles using peak heights," *Statistical Applications in Genetics and Molecular Biology*, vol. 15, pp. 431 – 445, 2016.

[21] S. Vilsen, T. Tvedebrink, H. Mogensen, and N. Morling, "Statistical Modelling of Ion PGM HID STR 10-plex MPS Data," *Forensic Science International: Genetics*, 2017.

[22] R. Cowell, S. Lauritzen, and J. Mortera, "Probabilistic expert systems for handling artifacts in complex dna mixtures," *Forensic Science International: Genetics*, vol. 5, pp. 202–209, 2011.

[23] T. Graversen and S. Lauritzen, "Computational aspects of DNA mixture analysis," *Statistics and Computing*, vol. 25, pp. 527–541, 2015.

[24] S. Vilsen, T. Tvedebrink, P. Eriksen, C. Bøsting, C. Hussing, H. Mogensen, and N. Morling, "Stutter analysis of complex STR MPS data," *Forensic Science International: Genetics*, vol. 35, pp. 107–112, 2018.

[25] C. Hussing, C. Huber, R. Bytyci, H. Mogensen, N. Morling, and C. Børsting, "Sequencing of 231 forensic genetic markers using the Illumina® ForeSeq™ workflow - an evaluation of the assay and software," *Forensic Sci. Res.*, 2018, (in press).

[26] C. Hussing, C. Huber, R. Bytyci, , N. Morling, and C. Børsting, "The Danish STR sequence database: Duplicate typing of 363 Danes with the ForenSeq™ DNA Signature Prep Kit," *Int. J. Legal Med.*, 2018, (in press).

[27] A. E. Woerner, J. L. King, and B. Budowle, "Fast STR allele identification with STRait Razor 3.0," *Forensic Science International: Genetics*, vol. 30, pp. 18–23, 2017.

[28] G. W. Brier, "Verification of forecasts expressed in terms of proability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1950.

[29] Riccardo Benedetti, "Scoring Rules for Forecast Verification," *Monthly Weather Review*, vol. 138, pp. 203–211, 2010.

Paper D

DNA mixture deconvolution using an evolutionary algorithm with multiple populations, hill-climbing, and guided mutation

Søren B. Vilsen, Torben Tvedebrink, and Poul Svante Eriksen

The paper is undergoing revisions.

The layout has been revised.

Abstract

DNA mixtures frequently occur in crime cases analysed in forensic genetics. These are convolutions of the DNA profiles of each contributor to the mixture. We introduce a multiple population evolutionary algorithm (MEA) used for the deconvolution of the DNA mixtures. The mutation operator of the MEA utilised that the fitness is based on a probabilistic model and was guided by the deviation between the observed and the expected value for every element of the encoded individual. The guided mutation operator (GMO) was designed such that the larger the deviation the higher probability of mutation. Furthermore, the GMO was inhomogeneous in time, decreasing to a specified lower bound as the number of iterations increased. This ensured that the operator would not fixate on the elements with large deviance residual, which could not be improved.

We analysed 30 two-person DNA mixtures in varying mixture proportions. The DNA profiles deconvoluted by the MEA were compared to the true DNA profiles, for a series of sensitivity experiments varying the subpopulation size, comparing a completely random homogeneous mutation operator to the guided operator with varying mutation decay rates, and allowing for hill-climbing of the parent individuals. We found that using either hill-climbing or GMO yield high quality solutions to our problem, while using RMO and no hill-climbing yield solutions of poorer quality and using both hill-climbing and GMO did not improve the fitness proportionally to added computational burden.

1 Introduction

Evolutionary algorithms, and meta-heuristics in general, have been shown to be very versatile tools for discrete and continuous optimisation in a wide range of applications [1–4]. This versatility is that they by nature are extremely simple yet flexible. Furthermore, they provide convergence to a global optimum, without requiring gradient or Hessian information to search through the space of possible solutions. This is achieved by repeatedly applying operators of Darwinian evolution, i.e. cross-over, mutation, and selection, which over time will breed better and better possible solution of the problem.

In most applications, the mutation operator is applied completely at random. Either by setting a flat mutation rate (which may be inhomogeneous in time) for every element of an individual, or by choosing the element(s) to be mutated at random. We will refer to this as the random mutation operator (RMO). The main advantage of the RMO is that it searches the state space thoroughly. However, if the state space is large relative to the set of optimal solutions, then most of the steps taken by the RMO will have little to no change of increasing the fitness of the individual (i.e. the proposed solution).

We propose to guide the mutation, called the guided mutation operator (GMO), by augmenting the probability of an element (of an individual) mutating based on the deviation seen between the observed data and what is expected given the decoded individual. Thus, the GMO was designed to have a larger chance of making mutation, which will increase the fitness of the individual, at the cost of not searching the state space as thoroughly as the RMO, and being more quickly fixated at a local optimum. In order to compensate for these two disadvantages, we added multiple randomly initialised sub-populations, and allowed for migration between these sub-population [5, 6].

The manuscript is organised as follows: Section 2 introduces the necessary background to understand the MEA implementation. In Section 3, we introduced the general structure of the MEA algorithm and its operators. Section 4 contains a sensitivity study and an examination of performance of the MEA implementation. Lastly, concluding remarks are given in Section 5.

2 Background

A central question when DNA evidence, $\mathcal{E}$, is presented in court is the determination of the posterior odds of the evidence under the competing hypotheses of the prosecution, $\mathcal{H}_p$, and defence, $\mathcal{H}_d$. Given DNA evidence, a DNA profile can be created, even from small samples of a few hundred pico-grams. A DNA profile is created by examining locations of the DNA, called markers. The markers are chosen such that they have a large variation in the population, but have a low mutation rate. At each (autosomal) marker a person may take two values, called alleles, and the collection of these alleles constitutes the persons DNA profile. If the sample is in large quantity and contains DNA from a single contributor, then identification is simple. However, when the DNA is found at a crime scene, the sample can be contaminated, be in extremely low quantity, contain DNA from multiple contributors, or any combination thereof [7, 8]. If the sample contains DNA from multiple contributors, called a DNA mixture, then accurately representing the posterior odds is very difficult, because the number of contributors, their relative contribution to the mixture, their DNA profiles, etc., are all unknowns.

In general the posterior odds of the two hypotheses can, by applying Bayes' theorem and denoting a probability by $\mathbb{P}(\cdot)$, be written as:

$$\frac{\mathbb{P}(\mathcal{H}_d|\mathcal{E})}{\mathbb{P}(\mathcal{H}_p|\mathcal{E})} = \frac{\mathbb{P}(\mathcal{E}|\mathcal{H}_d)}{\mathbb{P}(\mathcal{E}|\mathcal{H}_p)} \frac{\mathbb{P}(\mathcal{H}_d)}{\mathbb{P}(\mathcal{H}_p)}.$$

The prior odds of the two hypotheses should be supplied by the court, as it should represent the odds of the two hypotheses based on the evidence introduced to the court prior to the introduction of the DNA evidence. This leaves the likelihood ratio:

$$\mathrm{LR}(\mathcal{H}_d, \mathcal{H}_p) = \frac{\mathbb{P}(\mathcal{E}|\mathcal{H}_d)}{\mathbb{P}(\mathcal{E}|\mathcal{H}_p)},$$

for which, we write LR in the remainder of the manuscript.

The prevailing method of analysing DNA evidence is by determining the length of short tandem repeat (STR) regions using capillary electrophoresis (CE). In recent years, massively parallel sequences (MPS) has started to be introduced in forensic genetics casework. MPS

offers not just the length, but the entire base composition of the STR regions. That is, DNA samples analysed with MPS will be of higher resolution than those analysed with CE [9].

When a DNA sample is analysed it will yield some quantitative information $\boldsymbol{y}$ (called the coverage in the MPS setting) about some combined genetic information $\boldsymbol{g}_c$, structurally specified by a hypothesis $\mathcal{H}_i$. Therefore, we can factorise $\mathbb{P}\left(\mathcal{E}|\mathcal{H}_i\right)$ as:

$$\mathbb{P}\left(\mathcal{E}|\mathcal{H}_i\right) = \mathbb{P}\left(\boldsymbol{y}, \boldsymbol{g}_c|\mathcal{H}_i\right) = \mathbb{P}\left(\boldsymbol{y}|\boldsymbol{g}_c\right)\mathbb{P}\left(\boldsymbol{g}_c|\mathcal{H}_i\right).$$

That is, the probability of the evidence, given a hypothesis, has two parts: (1) The probability of the quantitative information given the genetic information [10], and (2) the probability of the genetic information under the hypothesis, which is usually assumed to follow a Dirichlet-Multinomial distribution, as described in Balding and Nichols (1994) [11].

It is suppressed that the probability of the quantitative information given the genetic information, $\mathbb{P}\left(\boldsymbol{y}|\boldsymbol{g}_c\right)$, will depend on unknown parameters, $\boldsymbol{\theta}$, describing the uncertainty of the measuring process, the signal intensities and the mixture proportions.

If a hypothesis, $\mathcal{H}_i$, states that a contributor is unknown it is necessary to sum over the set of possible unknown contributors, $\mathcal{U}$, to obtain the probability of the evidence. That is, if we assume (1) the sample contains DNA from two contributors, (2) one DNA profile is known, $\boldsymbol{g}_k$ (this could e.g. be the victim), and (3) $\mathcal{H}_i$ states that the DNA is a mixture of the known profile and a random unknown profile from the population. The evidence under the hypothesis is $\mathcal{E} = (\boldsymbol{y}, \boldsymbol{g}_k)$ and probability of the evidence reduces to:

$$\mathbb{P}\left(\boldsymbol{y}, \boldsymbol{g}_k|\mathcal{H}_i\right) \approx \sum_{\boldsymbol{g} \in \mathcal{U}} \mathbb{P}\left(\boldsymbol{y}|\boldsymbol{g}_k, \boldsymbol{g}, \hat{\boldsymbol{\theta}}_{\mathcal{U}}\right)\mathbb{P}\left(\boldsymbol{g}|\boldsymbol{g}_k\right), \tag{D.1}$$

where $\hat{\boldsymbol{\theta}}_{\mathcal{U}}$ is the $\boldsymbol{\theta}$ maximising the sum of the probabilities: $\sum_{\boldsymbol{g} \in \mathcal{U}} \mathbb{P}\left(\boldsymbol{y}|\boldsymbol{g}_k, \boldsymbol{g}, \boldsymbol{\theta}\right)$. The notation $\hat{\boldsymbol{\theta}}_{\mathcal{U}}$ should be interpreted as: the parameters were dependent on the entirety of the evidence, i.e. including every possible unknown genotype. Note that this formulation of the probability of the evidence can be extended to an arbitrary number of known and unknown DNA profiles.

The sum over the set of unknown contributors, $\mathcal{U}$, may be intractable because of the size of $\mathcal{U}$. In the CE setting, this problem

is largely solved by using a Bayesian network [12, 13], by sampling from the posterior distribution using Markov chains [14], or by simply limiting the number of unknown contributors to cases where the sum is tractable [15, 16]. However, as MPS is still relatively immature in the forensic genetics setting, the methods used for analysing CE DNA mixtures cannot be applied directly to MPS data, at least not without modification [10].

The set $\mathcal{U}$ is discrete and does not have any natural ordering, making an Evolutionary Algorithm (EA) the perfect tool. The algorithm presented here was implemented to achieve the following two objectives: (1) Find the combination of unknown genotypes maximising $\mathbb{P}\left(\boldsymbol{y}|\boldsymbol{g}_k, \boldsymbol{g}, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}\right)\mathbb{P}\left(\boldsymbol{g}|\boldsymbol{g}_k\right)$ w.r.t. $\boldsymbol{g}$, where the subscript in $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ indicates that the estimated parameters only depended on the unknown profile $\boldsymbol{g}$ and not the entire set $\mathcal{U}$, and (2) to approximate the set of combinations of unknown genotypes by the N_{top} unique individuals of the largest fitness seen throughout a run of the MEA.

The first objective is of interest, as finding the optimal unknown profile combination could be a useful place to start searching in a DNA database for a potential suspect in cases where the investigators had no other leads. The second objective, will be useful in stating the LR in cases with unknown contributors. Approximating the set of unknown contributor combinations, will make Eq. (D.1) tractable. The MEA described below was implemented in `R` and `C++` through the `Rcpp`-packages using the Eigen, Boost, and NLopt libraries [17–24] in the `R`-package MPSMixtures [25].

3 The multiple population evolutionary algorithm

We chose an MEA for the following three reasons: (1) the populations can be run in parallel, (2) if we initialise these populations randomly, then they will more thoroughly explore the fitness landscape, when compared to running a single populations algorithm, and (3) it allows us to utilise a smaller total population size, thereby, decreasing runtime. This implementation is a variation of the parallel evolutionary algorithm presented by Mühlenbein et al. [5, 6].

An outline of the implemented MEA can be seen in Algorithm 2. The MEA works by segregating the total population, $\mathcal{P}$, into $N_{\mathcal{P}}$ smaller

sub-populations, $\boldsymbol{P}_n$, each containing N_I individuals. The i'th individual of sub-population n will be referred to $\boldsymbol{p}_{ni}$, dropping subscripts if they are unnecessary or clear from context.

A single iterations of the MEA, consists of two phases:

(1) the migration phase (Algorithm 2: Line 4).

(2) the evolution phase (Algorithm 2: Lines 6-16).

During the migration phase, the $N_{\mathcal{P}}$ sub-populations exchange information according to a predefined pattern. The information exchanged and the migration pattern is described in Subsection 3.1 below. During the second phase, an EA is run on each sub-population independently, by in turn hill-climbing each individual in the population creating a parent, find a partner to the parent, use crossover to create a child of the individuals, mutate the child, and determine whether the child should replace its parent. A more detailed description of the selection, representation, and operators of the independent EAs can be found in Subsections 3.3 and 3.4 below. Note that in the entirety of this manuscript the word individual will always be used in the EA context, i.e. an individual can describe the genetic profile of multiple contributors (persons), whereas as person will always be referred to as a contributor.

The implemented MEA is said to have converged when the difference between the individuals of largest fitness of each sub-population is smaller than some ε for more than N_ε iterations.

Lastly, to keep track of the N_{top} unique individuals of largest fitness, we created and maintained the list $\boldsymbol{P}_{\text{top}}$ throughout a run of the MEA.

3.1 Migration

When migration occurs the highest fitness individual of the sub-population is copied and send to its neighbours, replacing the neighbours individual of lowest fitness. The migration operator, combined with the fact that the sub-populations were randomly initialised, creates the following advantage: If a sub-population gets stuck at a local maximum, then a migration of the highest fitness individual from another sub-population can help drag it out of the local maximum (hopefully

Algorithm 2 Multiple Population Evolutionary Algorithm.

Input: $\boldsymbol{y}$, $\boldsymbol{g}_k$, ε, $N_{\mathcal{P}}$, N_I, N_ε, N_{top}
Output: $\boldsymbol{P}_{\text{top}}$

$\mathcal{P}$: randomly initialise $N_{\mathcal{P}}$ sub-populations each with N_I individuals.
$\boldsymbol{P}_{\text{top}}$: initialised as an empty list.
while (not converged) **do**
  $\mathcal{P} \leftarrow \texttt{migrate}(\mathcal{P})$
  for (n from 1 to $N_{\mathcal{P}}$) **do**
    $\tilde{\boldsymbol{P}}$: empty sub-population.
    for (i from 1 to N_I) **do**
      $\tilde{\boldsymbol{p}} \leftarrow \texttt{hillclimb}(\boldsymbol{p}_{ni})$
      $\boldsymbol{q} \leftarrow \texttt{selectpartner}(\tilde{\boldsymbol{p}}, \boldsymbol{P}_n)$
      $\boldsymbol{c} \leftarrow \texttt{crossover}(\tilde{\boldsymbol{p}}, \boldsymbol{q})$.
      $\boldsymbol{c} \leftarrow \texttt{mutate}(\boldsymbol{c})$
      if ($F(\tilde{\boldsymbol{p}}) < F(\boldsymbol{c})$) **then**
        $\tilde{\boldsymbol{p}} \leftarrow \boldsymbol{c}$
      end if
      Append $\tilde{\boldsymbol{p}}$ to $\tilde{\boldsymbol{P}}$
    end for
    $\boldsymbol{P}_n \leftarrow \tilde{\boldsymbol{P}}$
  end for
  Update $\boldsymbol{P}_{\text{top}}$
  Update convergence criteria
end while
return $\boldsymbol{P}_{\text{top}}$

towards the global maximum, or at the very least push it towards another part of the sample space).

When designing the migration operator, we needed to balance the sharing of high fitness individuals, while simultaneously ensuring that the high fitness individuals did not spread too quickly, as the algorithm may then fixate at a local maximum. Therefore, we used the neighbourhood structure shown in Figure D.1 (inspired by [5] and [6]).

Using this structure and assuming that an individual of higher fitness was not created during this period, the minimum number of iterations needed for an individual to spread to every sub-population is exactly:

$$\lceil (N_{\mathcal{P}} + 1)/3 \rceil, \tag{D.2}$$

which follows by counting the number of ways to get to all states in both the clockwise and anti-clockwise direction.

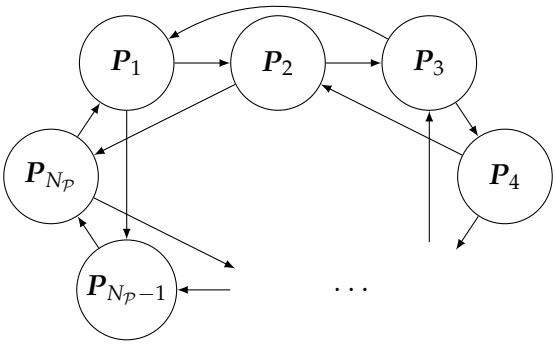

Fig. D.1: The migration neighbourhood structure used in the MEA.

3.2 Solution representation and fitness

3.2.1 The set of unknown genotype combinations

Any genotype on an autosomal marker will have exactly two alleles; they can be different or equal, called hetero- and homozygous, respectively, but there will always be exactly two alleles (disregarding extremely rare events). Therefore, any genotype on a marker m with A_m different observed alleles, denoted $\boldsymbol{g}_m$, can be represented as a vector with elements $g_{mi} \in \{0, 1, 2\}^{A_m}$, with $\sum_i g_{mi} = 2$.

In order to understand the shape and size of $\mathcal{U}$ and the solution representation used in the EA, we start by giving a small example.

Example 3.1 (Set of genotypes)
Assume we have a single marker with three observed alleles and a single unknown contributor, u. The possible unknown genotypes of u are:

$$\mathcal{U}_1 = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \right\},$$

i.e. a total of six possible genotypes. Note the subscript in $\mathcal{U}_1$ refers to the number of unknown contributors.

In general, assuming the total number of observations on a marker m is A_m, then the number of heterozgote and homozygote genotypes can be written as:

$$\binom{A_m}{2} \text{ and } A_m,$$

respectively. Thus, for M independent markers, the size of $\mathcal{U}_1$ is given as follows:

$$|\,\mathcal{U}_1\,| = \prod_{m=1}^{M} \frac{A_m(A_m+1)}{2}.$$

Furthermore, if there are U unknown contributors, then the size of $\mathcal{U}_U$ is

$$|\,\mathcal{U}_U\,| = |\,\mathcal{U}_1\,|^U.$$

However, as the order in which the genotypes appear will not affect the fitness, the size of the set can ultimately be reduced to:

$$|\,\mathcal{U}_U\,| = \binom{|\,\mathcal{U}_1\,| + U - 1}{U}.$$

3.2.2 Solution representation of unknown genotypes

We have encoded the genotypes of an unknown contributor using two elements per marker. This representation has two main advantages: (1) it simplifies the cross-over and mutation operators, and (2) it will always require less memory.

An individual will be encoded as pointers to the non-zero elements of the genotype matrix. Because the markers are assumed to be independent, encoding will be performed on a marker-by-marker basis, taking one contributor at a time.

Example 3.2 (Encoded individual)
Continued from Example 1, but assuming we have two unknown contributors. Furthermore, assume that the genotypes of the two unknown contributors are given by the matrix:

$$\boldsymbol{g} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{D.3}$$

where the first and second columns correspond to the first and second unknown contributor, respectively. Then, the encoded individual, $\boldsymbol{p}$, of $\boldsymbol{g}$ is:

$$\boldsymbol{p} = \begin{bmatrix} 1 \\ 3 \\ 1 \\ 2 \end{bmatrix}, \tag{D.4}$$

where the first two elements correspond to the first contributor and the last two elements to the second, respectively.

In general, when using the formulation seen in Eq. (D.4), it is going to require exactly 2 UM elements to store an individual, with U unknown contributors and M markers. Compared to the NU elements needed, if used the genotype matrix representation, seen in Eq. (D.3), where N is the number of observations in the dataset, i.e. $M = \sum_{m=1}^{M} A_m$.

3.2.3 Fitness of individuals

We are trying to find unknown genotype combination which maximises the probability of the evidence, i.e.

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g} \in \mathcal{U}} \left\{ \mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_k, \boldsymbol{g}, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}\right) \mathbb{P}\left(\boldsymbol{g} | \boldsymbol{g}_k\right) \right\},$$

where $\boldsymbol{g}_k$ are the known genotypes of the mixture, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is the estimated parameters given $\boldsymbol{g}$ (as well as $\boldsymbol{g}_k$ and $\boldsymbol{y}$). Therefore, we defined the fitness of an individual $\boldsymbol{p}$, corresponding to the unknown genotype(s) $\boldsymbol{g}$, as:

$$F(\boldsymbol{p}; \hat{\boldsymbol{\theta}}, \boldsymbol{y}, \boldsymbol{g}_k) = \mathbb{P}\left(\boldsymbol{y} | \boldsymbol{g}_k, \boldsymbol{g}, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}\right) \mathbb{P}\left(\boldsymbol{g} | \boldsymbol{g}_k\right), \tag{D.5}$$

Because of the definition of the fitness, it follows that any time an individual is changed, we will need to re-estimate the $\boldsymbol{\theta}$ parameters. For notational convenience, we write $F(\boldsymbol{p})$ instead of $F(\boldsymbol{p}; \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}, \boldsymbol{y}, \boldsymbol{g}_k)$ from this point forward.

3.3 Selection

Selection is split into two types: parent and survivor selection. The implemented survivor selection is extremely elitist. In order for a child to replace its parent, the fitness of the child has to be larger than the fitness of the parent otherwise the parent survives to the next iteration.

Parent selection in this implementation works slightly differently than in most EAs, as every individual gets to be a parent and a partner is then selected for that parent. With the partner being selected proportionally to its fitness. Furthermore, to avoid the population fixating a single solution too quickly, we restricted the search of the partner to a neighbourhood around the parent. That is, the probability of an individual $\boldsymbol{p}_j$ being selected as a partner of the parent $\boldsymbol{p}_i$ is defined as:

$$\pi^{(s)}(\boldsymbol{p}_j | \boldsymbol{p}_i) = \frac{F(\boldsymbol{p}_j)}{\sum_{k \in \mathcal{N}(\boldsymbol{p}_i, L)} F(\boldsymbol{p}_k)}, \tag{D.6}$$

for all $j \in \mathcal{N}(\boldsymbol{p}_i, L)$ and zero otherwise, where $\mathcal{N}(\boldsymbol{p}_i, L)$ is the neighbourhood of $\boldsymbol{p}_i$ defined as:

$$\mathcal{N}(\boldsymbol{p}_i, L) = \left\{ (i + l) \bmod N_I \right\}_{l \in [-L; L] \setminus \{0\}}, \tag{D.7}$$

i.e. a window of size $2L$ with i as its midpoint.

The size of the neighbourhood can be used to control the time to fixation on the sub-population level; the smaller the neighbourhood the longer the time to fixation.

3.4 Operators

3.4.1 Crossover

Because of the nature of the implemented parent and survivor selection, we have chosen a crossover operator which creates a single child. The procedure is sketched in Algorithm 3. Given two individuals, a parent and its partner, the child is created by copying elements one-by-one from either the parent of the partner (the algorithm always starts with parent). After an element has been copied, we switch from parent to partner (or vice versa) with probability of $\pi^{(c)}$, creating a single child of the two individuals. The probability of switching will be inversely proportional with the length of the individual, i.e. $\pi^{(c)} \propto 1/(2\ UM)$. That is, any newly created child will have experienced a single crossover event on average independent of length.

Algorithm 3 Crossover

Input: p, q, $\pi^{(c)}$
Output: c

$s =$ `false`
for $(i$ from 0 to $(2\ UM - 1))$ **do**
  $u \sim \text{Unif}(0, 1)$.
  if $\left(u < \pi^{(c)}\right)$ **then**
    $s = \neg\, s$
  end if
  if (s) **then**
    $c_i \leftarrow p_i$
  else
    $c_i \leftarrow q_i$
  end if
end for
return c

3.4.2 Mutation

The mutation operator works on an element-by-element basis choosing to mutate element i with probability $\pi_i^{(m)}$.

If we choose to mutate an element c_i, then the element is changed by drawing a random number $a \in \{1, ..., A_m - 1\}$ and updating the element, as:

$$c_i = (c_i + a) \mod A_m \tag{D.8}$$

Note that if an element is mutated it always changes its value, as the '0' state is already included as $(1 - \pi_i^{(m)}) > 0$ for all i.

The guided nature of the mutation is controlled through the mutation probabilities. The pseudo-code of the implemented GMO is seen in Algorithm 4.

Under the probability model given $\boldsymbol{g}$, the observed data $\boldsymbol{y}$ will have an expected value, $\hat{\boldsymbol{y}}$. A standardised residual measures the deviation between these two vectors. Depending on the type of model there exists different types of standardised residual. In particular, the model used is a generalised linear model and we will, therefore, choose the so called deviance residuals, as they are approximately normally distributed. We denote the deviance residuals by r^D. With this in mind, we defined the probability of mutating in the t'th iteration of an independent EA, as:

$$\begin{aligned} &\pi_i^{(m)}\left(r_i, \pi_{\text{LB}}^{(m)}, \pi_{\text{UB}}^{(m)}\right) \\ &\quad = \pi_{\text{UB}}^{(m)} - \left(\pi_{\text{UB}}^{(m)} - \pi_{\text{LB}}^{(m)}\right) \frac{f(r_i)}{f(0)} \end{aligned} \tag{D.9}$$

where r_i are the deviance residuals, f is the density function of a standard normal distribution, and $\pi_{\text{LB}}^{(m)}$ and $\pi_{\text{UB}}^{(m)}$ are a lower and upper bound on probability of mutation, respectively. The lower and upper bounds were introduced to ensure that the transition matrix was (still) fully connected. That is, they were introduced to ensure that the method still converges towards a global maximum. We have shown a plot of the function in Fig. D.2 with $\pi_{\text{LB}}^{(m)} = 0.05$ and $\pi_{\text{UB}}^{(m)} = 0.95$.

The probability of mutation described in Eq. (D.9) is very useful in the beginning iterations of MEA, but becomes less and less effective as the number of iterations increases. As it will keep mutating the

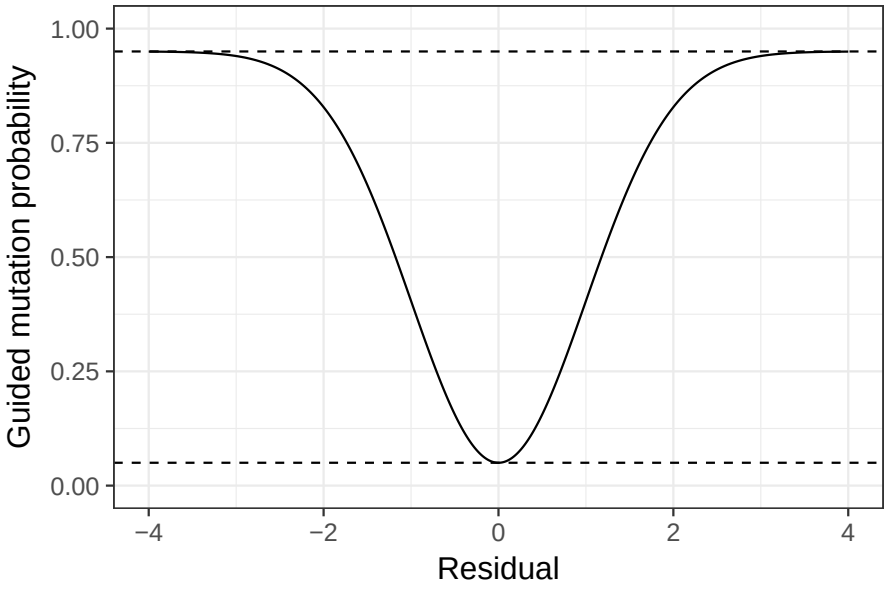

Fig. D.2: The guided mutation probability against the residual.

same elements again and again, even if the element is as good as its ever going to get. Therefore, we introduced an iteration dependent upper bound which will tend towards the lower bound as the number of iterations increased. In order to ease notation, we write $\pi^{(m)}_{\mathrm{UB},t}$ and $\pi^{(m)}_{i,t}$ instead of $\pi^{(m)}_{\mathrm{UB}}(t)$ and $\pi^{(m)}_{i}\left(r_i, \pi^{(m)}_{\mathrm{LB}}, \pi^{(m)}_{\mathrm{UB},t}\right)$, respectively. The behaviour of the resulting probability of mutation is depicted in Fig. D.3, using $\pi^{(m)}_{\mathrm{LB}} = 1 - \pi^{(m)}_{\mathrm{UB},0} = 0.05$. Note that $\left\{\pi^{(m)}_{\mathrm{UB},t}\right\}_{t\leq 0}$ can be taken as any non-increasing sequence, with initial value less than or equal to one. Thus, the rate of decay, x, can be specified as any strictly positive real number, i.e. $x \in \mathbb{R}^+$. Setting the decay rate at $x = 2$ implies that after $N_{\max}/2$ iterations the guided mutation probability is at $\pi^{(m)}_{\mathrm{LB}}$ for every element of the individual.

3.4.3 Hill-climbing

We chose to include hill-climbing of the parents in an effort to increase the average fitness of the population even if no of the mutated child reached a higher fitness than the parent. Furthermore, we hill-climb

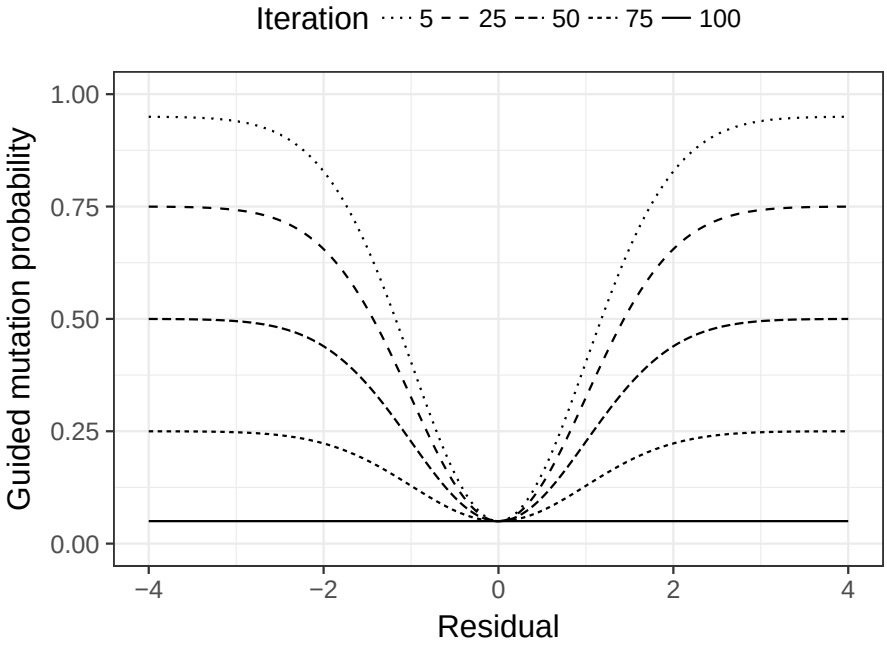

Fig. D.3: The guided mutation probability against the residual, using an upper bound inhomogeneous in time, hitting the lower bound after 100 iterations.

the parent before choosing its partner, thereby, giving the child the best possible outset before mutation.

The guided hill-climbing algorithm, seen in Algorithm 5, randomly chooses an element of the individual, creates the $(A_m - 1)$ remaining instances of that element and would ideally re-calculate the fitness for every instance. However, this requires re-estimating the parameters, $\boldsymbol{\theta}$, for every instance, which is by far the slowest part of the implementation. Therefore, we will take a slightly different approach using the raw residuals. The raw residuals are defined as the difference between the observed and expected vectors.

If the raw residuals are negative (positive), then the expected coverage is larger (smaller) than the observed coverage. If they were extremely negative it could imply that the algorithm was expecting a true allele, but is pointing to a what is most likely noise (reversed for extremely positive residuals). In this case, we would like the algorithm to change from pointing at the potential noise, to pointing at a true

Algorithm 4 Guided mutation operator (GMO)

Input: $\boldsymbol{c}$, $\pi_{\text{LB}}^{(m)}$, $\pi_{\text{UB}}^{(m)}(t)$
Output: $\boldsymbol{c}$

for $(i$ from 0 to $(2\,UM-1))$ **do**
  $m \leftarrow \left\lfloor \dfrac{i}{2\,UM} \right\rfloor$
  $a \leftarrow c_i$
  $r_i \leftarrow r^D\left(y_{ma}, \hat{\mu}_{ma}, \dfrac{\hat{\mu}_{ma}}{\hat{\gamma}}\right)$
  $\pi_{i,t}^{(m)} \leftarrow \left(1-\pi_{\text{UB},t}^{(m)}\right) - \left(1-\pi_{\text{UB},t}^{(m)}-\pi_{\text{LB}}^{(m)}\right)\dfrac{f(r_i)}{f(0)}$
  $u \sim \text{Unif}(0,1)$.
  if $\left(u < \pi_{i,t}^{(m)}\right)$ **then**
    $s \sim \text{Unif}\{1,2,...,A_m-1\}$
    $c_i \leftarrow (c_i+s) \bmod A_m$
  end if
end for
return $\boldsymbol{c}$

allele, without having to re-estimate the parameters.

Our solution was to choose the new instance, j, such that $r_i = -r_j$ (or $r_i + r_j = 0$). This would ensure that the squared residuals would not change, and, thus, the estimated parameters would not change. In cases where we could not find such an instance, we chose j such that the sum of residuals got as close to zero as possible, i.e.:

$$j = \arg\min_k |r_i + r_k|.$$

For this instance only, we estimated the parameters and compared the fitness of the new instance to that of the parent. If the fitness of the new instance is larger than that of the parent, then it replaced the parent.

4 Experiments and results

Algorithm 5 Guided hill-climbing

Input: $\boldsymbol{p}$
Output: $\boldsymbol{p}$

for $(h$ from 1 to $N_H)$ **do**
  $i \sim \text{Unif}\{0, 1, ..., 2\,UM - 1\}$
  $\mathcal{I}$: Empty list of size $A_m - 1$.
  for $(a$ from 1 to $A_m - 1)$ **do**
    $\boldsymbol{s} \leftarrow \boldsymbol{p}$
    $s_i \leftarrow (s_i + a) \bmod A_m$
    $\mathcal{I}[a] \leftarrow \boldsymbol{s}$
  end for
  $\boldsymbol{k} \leftarrow \underset{\boldsymbol{s} \in \mathcal{I}}{\arg\min} \left\{ \left| r^D(p_i) + r^D(s_i) \right| \right\}$
  if $\left(F(\boldsymbol{p}) < F(\boldsymbol{k})\right)$ **then**
    $\boldsymbol{p} \leftarrow \boldsymbol{k}$
  end if
end for
return $\boldsymbol{p}$

4.1 The data

The analysis presented below was based on 15 controlled two person mixture experiments sequenced in duplicate, resulting in 30 two-person DNA mixture samples. DNA from two contributors, one male and one female, were used to create the experiments in the following mixture ratios: 1000:1, 100:1, 50:1, 25:1, 12:1, 6:1, 3:1, 1:1, 1:3, 1:6, 1:12, 1:25, 1:50, 1:100, 1:1000.

The 30 samples were sequenced using the Illumina MiSeq and the ForenSeq Panel B kit [26]. We restricted the analysis to the autosomal STRs, i.e. a total of 27 markers were considered in the analysis. Furthermore, the true DNA profiles of both contributors were determined in a previous study [10, 26].

The experiments were created and sequenced at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

Table D.1: Comparing the quality of the solution and the time to convergence, assuming the minor profile was known, when varying the number of sub-populations while keeping the total population size fixed at 2,000. The quality of the solutions was quantified by: (1) the average percentage of matching alleles and markers between the optimal unknown major profile and the true major profile, and (2) the number of samples where the fitness of the optimal unknown major profile is greater than or equal to the fitness of the true major profile.

		Time (iterations)				Time (hours)				Number of times $F_{\text{optimal}} \geq F_{\text{true}}$			
	# Sub-populations	2	4	8	16	2	4	8	16	2	4	8	16
Major:Minor	1000:1	56	37	39	44	8.19	3.41	1.82	1.10	3	3	3	3
	100:1	36	37	41	45	5.77	3.84	2.11	1.19	4	4	4	4
	50:1	37	39	44	44	6.55	4.15	2.28	1.15	3	3	3	4
	25:1	37	37	39	44	8.18	4.85	2.42	1.51	4	4	4	4
	12:1	34	38	40	48	5.88	4	2.08	1.17	4	4	4	4
	6:1	34	38	39	44	5.23	3.16	1.81	1.06	4	4	3	4
	3:1	36	36	39	45	5.85	3.22	1.86	1.10	4	4	4	4
	1:1	37	39	46	47	6.69	3.97	2.48	1.26	2	2	2	2
	Total									28	28	27	29

4.2 Sensitivity study

We examined the effectiveness of the guided operators and the number of sub-populations used (fixing the total population size), on the deconvolution when assuming the profile of the contributor with smallest amount of DNA, called the minor contributor, was known. In particular, we conducted the following sensitivity experiments:

(1) Using the RMO with the number of hill-climbing iterations at 0 or 2.

(2) Using the GMO with the mutation decay rate equal to 1/2, 1, or 2 and the number of hill-climbing iterations at 0 or 2.

(3) Using 2, 4, 8, and 16 sub-populations, with the total population size fixed at 2,000.

In all thirteen cases, we measured the time it took to converge, in both iterations and hours, and compared the fitness of the individual with highest fitness, F_{optimal}, to the fitness of the true major profile, F_{true}, by counting the number of samples for which the F_{optimal} was

larger than or equal to F_{true}. Furthermore, we calculated the average percentage of matching alleles and markers between the optimal unknown major profile and the true major profile.

Table D.1 shows the results of the sub-population sensitivity experiment, Item (3) in the list above. We see that the quality of the solutions were roughly equal for all four sub-population sizes. However, we find the strength of increasing the number of sub-population is shown in the hours to convergence. As the number of sub-populations doubles the median hours to convergence is reduced by a factor between 1.6 and 1.8. The reason the factor was not exactly 2 was because the number of iterations to convergence increased with the number of sub-populations. This is a direct consequence of the sub-population structure, as the optimal solution would take longer to propagate through the sub-populations as the number of sub-populations increased. Furthermore, we should note that the number iterations to convergence did not increase as much as we would have expected looking at Eq. (D.2). This is likely due to the added diversity introduced by running more randomly initialised sub-populations.

The results of the sensitivity studies outlined in Items (1) and (2), are shown in the upper and lower part of Table D.2, respectively. Starting with the upper half, we see that the speed of the mutation decay does not have an effect on the quality of the solutions. However, the table shows that if we do not hill-climb, then the quality of the optimal solutions when using RMO (the 'NA' columns in the table) is much lower than when using GMO. We also see that the GMO added nearly no additional computational time. Looking at the lower part, we see that the quality of the optimal solutions is poor only when setting the mutation decay rate to 2. That is, if using hill-climbing the mutation does not need to be guided to obtain high quality solutions. Comparing the tables, we see that either the GMO or the hill-climbing should be used to achieve high quality optimal solution. Furthermore, we see that using hill-climbing significantly increased the time to convergence, without increasing the quality of the solutions.

Thus, we will from this point forward use the following settings: 16 sub-populations, 125 individuals per sub-populations, 10 inner iterations, 250 outer iterations, mutation decay rate of 1, and no hill-climbing.

Table D.2: Comparing the quality of the solution and the time to convergence, assuming the minor profile was known, when varying the number hill-climbing iterations, and the mutation decay rate. A mutation decay rate of 0 is used to indicate that the mutation was not guided, but completely random.

Hill-climbing iterations = 0

		Identical alleles (%)				Identical markers (%)				Time (hours)				Number of times $F_{optimal} \geq F_{true}$			
	Mutation decay	NA	1/2	1	2	NA	1/2	1	2	NA	1/2	1	2	NA	1/2	1	2
Major:Minor	1000:1	0.99	0.98	0.97	0.97	0.98	0.97	0.96	0.95	0.57	0.59	0.57	0.56	1	4	4	3
	100:1	0.97	0.96	1	0.99	0.96	0.93	0.99	0.98	0.59	0.58	0.66	0.64	2	4	4	4
	50:1	0.96	0.97	0.96	0.98	0.95	0.96	0.94	0.97	0.65	0.61	0.64	0.64	1	4	3	3
	25:1	0.96	0.96	0.98	0.95	0.93	0.94	0.96	0.93	0.62	0.82	0.86	0.80	2	4	4	4
	12:1	0.96	0.97	0.96	0.99	0.93	0.95	0.93	0.98	0.90	0.59	0.64	0.60	4	3	4	4
	6:1	0.98	0.99	1	0.99	0.97	0.98	0.99	0.98	0.56	0.56	0.58	0.54	4	4	4	4
	3:1	0.94	0.96	0.97	0.96	0.92	0.95	0.96	0.93	0.56	0.56	0.58	0.56	4	4	4	4
	1:1	0.91	0.94	0.94	0.92	0.84	0.88	0.88	0.88	0.69	0.79	0.70	0.72	2	2	2	2
	Total	0.96	0.97	0.97	0.97	0.94	0.94	0.95	0.95	0.90	0.82	0.86	0.80	20	29	29	28

Hill-climbing iterations = 2

		Identical alleles (%)				Identical markers (%)				Time (hours)				Number of times $F_{optimal} \geq F_{true}$			
	Mutation decay	NA	1/2	1	2	NA	1/2	1	2	NA	1/2	1	2	NA	1/2	1	2
Major:Minor	1000:1	0.99	0.97	0.99	0.99	0.97	0.94	0.98	0.98	1.27	1.77	1.29	1	3	3	3	1
	100:1	0.99	0.98	0.99	0.98	0.97	0.97	0.98	0.96	1.36	1.57	1.33	1.07	4	4	4	3
	50:1	0.99	0.98	0.98	0.98	0.98	0.96	0.97	0.96	1.28	2.35	1.59	1.15	4	3	4	3
	25:1	0.98	0.96	0.98	0.99	0.96	0.95	0.96	0.98	1.76	1.64	1.37	1.43	4	4	4	4
	12:1	0.98	0.96	0.98	0.97	0.96	0.95	0.96	0.95	2.57	1.20	1.11	1.18	4	4	4	4
	6:1	0.98	0.98	0.99	0.99	0.96	0.96	0.98	0.98	1.19	1.05	0.95	1.05	4	4	4	4
	3:1	0.96	0.97	0.96	0.96	0.94	0.94	0.94	0.94	1.21	1.10	0.99	1.51	4	4	4	4
	1:1	0.93	0.94	0.90	0.91	0.86	0.88	0.82	0.84	1.72	1.44	1.19	1.22	2	2	2	2
	Total	0.97	0.97	0.97	0.97	0.95	0.94	0.95	0.95	2.57	2.35	1.59	1.51	29	28	29	25

4.3 Deconvolution

For each of the 30 samples, we conducted two deconvolution experiments: (1) assuming the minor profile was known, (2) assuming the major profile was known. In both cases, we counted the number of alleles and the number of markers the optimal profile had in common with the true profile. The first case should be trivial, as the major profile should be clear on most marker and we expected the deconvoluted profile to be within a marker or two of the true profile, and it is used as a benchmark for the implemented MEA [25].

The average percentage is shown in Table D.3. We found that when the minor is known, we could correctly identify more than 94% of the alleles and above 88% of entire markers independently of the mixture ratio. When assuming the major is known, we saw that the number of correctly identified alleles and markers increased steadily as the

mixture proportions tends to an even 1:1 mixture ratio. Furthermore, it is worth noting that when the major is known, the fitness of the optimal individual was always larger than the fitness assuming both major and minor were known (with the exception of the 1:1 mixture samples). That is, the optimal minor profile always yielded a better fitting profile than the true minor profile (not shown here).

Table D.3: The percentage of identical alleles and markers, when comparing the true profile of the unknown contributor with the optimal profile found by the MEA. The MEA used 16 sub-populations with a size of 125, GMO with a decay of 1, and no hill-climbing.

		Identical alleles (%)		Identical markers (%)	
		Known major	Known minor	Known major	Known minor
Major:Minor	1000:1	0.23	0.97	0.06	0.95
	100:1	0.29	0.94	0.07	0.90
	50:1	0.31	0.97	0.13	0.96
	25:1	0.40	0.95	0.15	0.93
	12:1	0.60	1.00	0.40	1.00
	6:1	0.74	0.99	0.58	0.98
	3:1	0.86	0.96	0.75	0.93
	1:1	0.97	0.94	0.95	0.88

4.4 The set of unknown genotypes

We considered three approximations to the set of unknown genotypes, $\mathcal{U}$. Two of the approximations were based on the set of N_{top} fittest individuals, denoted $\mathcal{U}_{N_{\text{top}}}$, found by the MEA. The third approximation samples from the posterior distribution of the unknown genotypes, using the Metropolis-Hastings (MH) algorithm.

The approximations will be compared by examining the probability of the evidence (PoE), seen in Eq. (D.1), and comparing it to the exact PoE. In order to make the calculation of the exact PoE manageable, we considered only a single unknown contributor. Furthermore, we note that Eq. (D.1), can be written as

$$\prod_m \sum_{g_m \in \mathcal{U}_m} \mathbb{P}\left(\boldsymbol{y}_m | \boldsymbol{g}_{mk}, \boldsymbol{g}_m, \hat{\boldsymbol{\theta}}_{\mathcal{U}}\right) \mathbb{P}\left(\boldsymbol{g}_m | \boldsymbol{g}_{mk}\right), \tag{D.10}$$

because the markers were assumed conditionally independent, given the parameters. This formulation of the PoE simplifies both the calculation of the exact PoE and the sampling, as it can be performed on a marker-by-marker basis. However, the estimated parameters still depends on the entire set $\mathcal{U}$.

The MEA produces entire unknown genotypes and, thus, the formulation in Eq. (D.1) is more appropriate than the formulation introduced in Eq. (D.10). Therefore, the first approximation using the MEA, will directly employ the set $\mathcal{U}_{N_{\text{top}}}$ in Eq. (D.1). The second MEA approximation takes the unique genotypes on every marker of $\mathcal{U}_{N_{\text{top}}}$ to approximate $\mathcal{U}_m$ in Eq. (D.10).

As we were primarily interested in the approximation to the set $\mathcal{U}$, we will use the same parameter estimates in all four cases. We chose to use the parameters estimated using $\mathcal{U}_{N_{\text{top}}}$. This eliminates the variability introduced by re-estimating the parameters in each of the four cases, thereby, making the comparison as fair to the MEA as possible.

In all three cases, the approximations of the PoE, $\hat{p}(\mathcal{E})$, were compared to the exact PoE, $p(\mathcal{E})$, by the absolute relative difference, i.e.

$$\left| \frac{p(\mathcal{E}) - \hat{p}(\mathcal{E})}{p(\mathcal{E})} \right|.$$

The MH approximation took 10,000 steps for each marker, totalling 280,000 iterations, starting at the individual with the highest fitness, in the list of fittest individuals returned by the MEA. Starting in the optimal individual we eliminated the need for a burn-in, as we are expected to start the sampling in a high-posterior region.

Figure D.4 shows the relative difference between each of the three approximations and the exact PoE. We see that the MH method provides by the best approximation of the three approaches. This is to be expected as the number of profiles sampled ($N = 280{,}000$) in the MH method by far outweighs the number of profiles stored by the MEA ($N = 1{,}000$). Regarding the two types of MEA approximation, we see that relative difference of the second MEA approximation is always smaller than or equal to the first MEA approximation. In particular, we note that the relative difference of the second MEA approximation never exceeds 0.5%, while the first approximation stays below 3% (in most cases below 2.5%), when compared to the exact PoE.

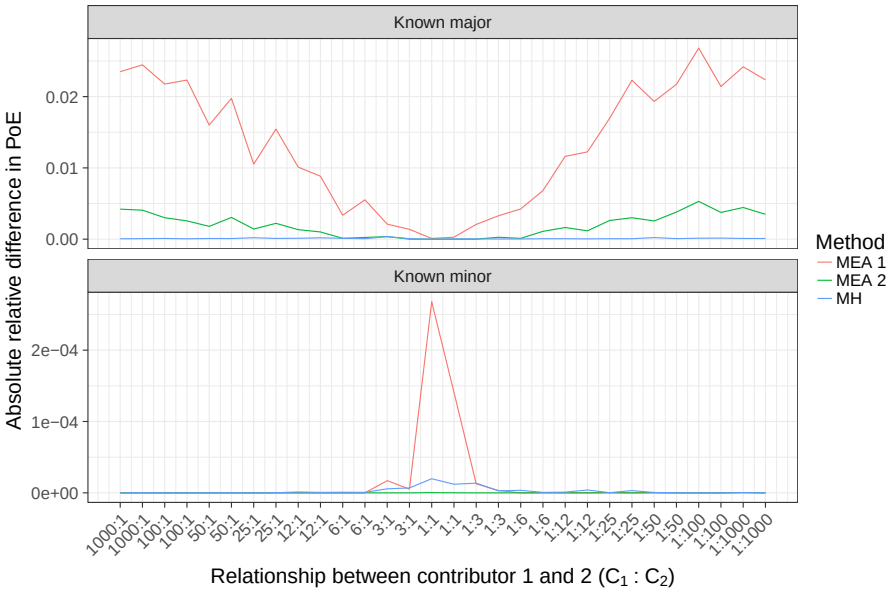

Fig. D.4: The relative difference for each sample. The abscissa axis is labelled according to the mixture ratio of the two contributors (C_1 and C_2) to the sample (shown as C_1:C_2). The relative difference is shown for all three approximation methods, and assuming that either the minor or major contributors profile was known. MEA 1 and 2 represents the approximations using $\mathcal{U}_{N_{\text{top}}}$ directly and using the unique genotypes on a marker-by-marker basis, respectively. MH is the Metropolis-Hastings sampler, sampling genotypes on a marker-by-marker basis.

Lastly, we see that if minor profile was known then there was no difference between the MEA and MH approximations, as the optimal profile will tend to dominate the sum.

5 Concluding remarks

The implemented MEA takes, for a single unknown contributor, approximately an hour to converge when not using any hill-climbing, with hill-climbing this increases to approximately two hours, as seen in Tables D.1-D.3. This increase in computational time is a consequence of needing to re-estimate the parameters in every hill-climbing iteration. Thus, when using hill-climbing the MEA needed to estimate the parameters an additional two times for every individual in every population in every inner iteration of the algorithm. Because of the model,

the parameter estimation is a slow process which takes between 100 and 180 milliseconds, dependent of the encoded individual. Thus, we could achieve a large decrease in computational time by reducing the time it take to estimate the parameters. Whether this is achieved by using a different optimisation routine, or by approximating the parameter estimates and/or the likelihood, we have left to future research.

It should also be noted that the size of the sub-populations was higher than strictly necessary, which is why the results were fairly consistent between the methods. Decreasing the number of individuals in every sub-populations would also greatly decrease the computational time of the algorithm. Furthermore, we saw no real consequence in increasing the number of sub-populations. It could be interesting to run a wider range of sub-populations to see if this holds true as long as we have enough free arithmetic cores available, or if we hit a plateau at some point.

From our sensitivity studies it would seem that as long as we have something guiding the evolution of the individuals, then the algorithm returned high quality solution to our problem. However, using both hill-climbing and GMO was a waste of computational resources. This could also be a consequence of the relatively large total population size. That is, using both hill-climbing and GMO may still be beneficial for smaller population sizes.

Lastly, from section 4.4, it should be clear that the approximation of the PoE should be made using the MH sampler. However, that does not imply that the list of fittest individuals found by the MEA, denoted the N_{top}-list, is not useful, as the implemented MH sampler depends on the parameters estimated using the N_{top}-list. Thus, N_{top} should still be larger than one, even when not used to approximate the PoE. Furthermore, it should be noted, that if the parameters were not estimated and assumed known, then the MH sampling would become a lot more complicated, as the markers can no longer be sampled independently.

Acknowledgment

The authors would like to thank associate professor Leif K. Jørgensen for his help in deriving the minimum number of iterations needed for an individual to spread to every sub-population.

References

[1] C. Blum and A. Roli, "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison," *ACM Computing Surveys*, vol. 35, pp. 268–308, 2003.

[2] Z. Tu and Y. Lu, "A Robust Stochastic Genetic Algorithm (StGA) for Global Numerical Optimization," *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, vol. 8, pp. 456–470, 2004.

[3] S. Das and P. N. Suganthan, "Differential Evolution: A Survey of the State-of-the-Art," *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, vol. 15, pp. 4–31, 2011.

[4] A. Gogna and A. Tayal, "Metaheuristics: review and application," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 25, pp. 503–526, 2013.

[5] H. Mühlenbein, "Evolution in Time and Space - The Parallel Genetic Algorithm," *Foundations of Genetic Algorithms*, pp. 316 – 338, 1991.

[6] H. Mühlenbein, M. Schomisch, and J. Born, "The parallel genetic algorithm as function optimizer," *Parallel Computing*, pp. 619 – 632, 1991.

[7] J. Butler, *Fundamentals of Forensic DNA Typing*. Academic Press, 2009.

[8] ——, *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, 2012.

[9] C. B. rsting and N. Morling, "Next generation sequencing and its applications in forensic genetics," *Forensic Science International: Genetics*, vol. 18, pp. 78 – 89, 2015.

[10] S. Vilsen, T. Tvedebrink, P. S. Eriksen, C. Hussing, C. B. rsting, and N. Morling, "Modelling allelic drop-outs in STR sequencing data generated by MPS," *Forensic Science International: Genetics*, 2018.

[11] D. Balding and R. Nichols, "DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands," *Forensic Science International*, vol. 64, pp. 125 – 140, 1994.

[12] R. Cowell, T. Graversen, S. Lauritzen, and J. Mortera, "Analysis of Forensic DNA Mixtures with Artefacts," *Royal Statistical Society. Journal Series C: Applied Statistics*, vol. 64, pp. 1 – 32, 2015.

[13] Ø. Bleka, G. Storvik, and P. Gill, "EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts," *Forensic Science International: Genetics*, vol. 21, pp. 35 – 44, 2016.

[14] D. Taylor, J.-A. Bright, and J. Buckleton, "The interpretation of single source and mixed DNA profile," *Forensic Science International: Genetics*, vol. 7, pp. 516 – 528, 2013.

[15] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling, "Identifying contributors of DNA mixtures by means of quantitative information of STR typing," *Journal of Computational Biology*, vol. 19, pp. 887 – 902, 2012.

[16] C. Steele, M. Greenhalgh, and D. Balding, "Evaluation of low-template DNA profiles using peak heights," *Statistical Applications in Genetics and Molecular Biology*, vol. 15, pp. 431 – 445, 2016.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[18] B. Stroustrup, *The C++ Programming Language*, 4th ed. Addison-Wesley Professional, 2013.

[19] D. Eddelbuettel and R. François, "Rcpp: Seamless R and C++ integration," *Journal of Statistical Software*, vol. 40, no. 8, pp. 1–18, 2011. [Online]. Available: http://www.jstatsoft.org/v40/i08/

[20] G. Guennebaud, B. Jacob *et al.*, "Eigen v3," http://eigen.tuxfamily.org, 2010.

[21] D. Bates and D. Eddelbuettel, "Fast and elegant numerical linear algebra using the RcppEigen package," *Journal of Statistical Software*, vol. 52, no. 5, pp. 1–24, 2013. [Online]. Available: http://www.jstatsoft.org/v52/i05/

[22] B. Schling, *The Boost C++ Libraries*. XML Press, 2011.

[23] D. Eddelbuettel, J. W. Emerson, and M. J. Kane, *BH: Boost C++ Header Files*, 2018, R package version 1.66.0-1. [Online]. Available: https://CRAN.R-project.org/package=BH

[24] S. G. Johnson, "The NLopt nonlinear-optimization package," http://ab-initio.mit.edu/nlopt, 2010.

[25] S. ren B. Vilsen, "The MPS mixtures package," https://github.com/svilsen/MPSMixtures, 2018.

[26] C. Hussing, C. Huber, R. Bytyci, , N. Morling, and C. Børsting, "The Danish STR sequence database: Duplicate typing of 363 Danes with the ForenSeq^{TM} DNA Signature Prep Kit," *Int. J. Legal Med.*, 2018, (in press).

Paper E

Analysing MPS STR DNA mixture samples

Søren B. Vilsen, Torben Tvedebrink, Poul Svante Eriksen, Claus Børsting, Christian Hussing, and Niels Morling

The paper is undergoing revisions.

The layout has been revised.

Abstract

We updated the Poisson-gamma coverage model, presented in Vilsen et al. [1], to better analyse DNA mixture samples quantified by Massively Parallel Sequencing (MPS). The updated model accounts for an arbitrary level of stuttering, sample adaptable estimates of the marker imbalances, and a slight change to the variance of the model. The updated model was used to analyse both between and within sample normalisation, an approximation of the Poisson-gamma distribution, and the estimated mixture proportions of the major contributor to a series of two person mixture samples quantified using the ForenSeq^{TM} DNA Signature Prep Kit.

The relationship between the major and minor contributor was maintained, i.e. we observed no within sample normalisation. Furthermore, we found that the between sample normalisation used did not force the number reads of each sample to be equal, contrary to Vilsen et al. [2]. Instead an upper limit to the allowed amount of template DNA is set. This implied that the coverage of the samples would be proportional to the amount of template DNA used, as long as the amount of template DNA did not raise above the upper limit. We approximated the updated Poisson-gamma coverage model using a gamma distribution with equivalent mean and variance. We found the gamma distribution to be a good approximation as long as the amount of template DNA was larger than 62.5 pg (an average coverage of approximately of 25). In samples with less than 62.5 pg template DNA, we start observing allele drop-outs. These are naturally handled by the Poisson-gamma model, but not the gamma model. Lastly, we saw that the average relative difference between the estimated and true mixture proportion of the major contributor was less than 2%.

1 Introduction

When DNA is found at a crime scene it may be in low quantity, degraded, contain DNA from more than one contributor, or some combination thereof. If the sample contain DNA from more than a single contributor it is referred to as a DNA mixture. The prevailing technique for quantifying short tandem repeat (STR) regions is capillary electrophoresis (CE) [3, 4]. The CE quantification yields information about the amount of DNA fragments of a given length found in the sample. However, with the introduction of Massively Parallel Sequencing (MPS) to forensic genetics casework [5–12], we can now obtain the base composition of the STR regions of the DNA sample. That is, MPS offers a higher resolution of the alleles than CE. This added resolution will be useful in determining the contributors to a mixture as well as their relative mixture proportions [12]. Thus, we want to model DNA mixtures quantified by MPS.

It is not possible to directly translate the methods utilised for DNA mixture samples quantified by CE to the MPS setting. However, as both the CE and the MPS processes are based on polymerase chain reaction (PCR) amplification, some of the lessons learned from CE models were useful in the development of the MPS coverage model [1, 13]. The MPS coverage model took the mean structure of the 'gamma model' [14–16] and extended it to account for the MPS process's marker imbalances and added resolution. The MPS coverage model modified the gamma model to: (1) account for the between marker variation (also called marker imbalances), (2) use a predictor of the rate of stuttering utilising the added resolution of the MPS process, and (3) change the choice of distribution.

A step in any MPS workflow is sample normalisation [2]. Sample normalisation is necessary, because more than one sample can be analysed on a chip in parallel. We do not want a small number of samples to completely dominate the chip/plate used for quantification. We distinguished between two types of sample normalisation: (1) ensuring that every sample quantified on the chip contained the same number of reads and (2) enforcing an upper bound on the number of reads for each sample on the chip. Of the two types, the latter is preferred as it may maintain the relationship between samples (if the amount of input DNA is smaller than the upper bound), which the former

does not. Lastly, if the reads are chosen at random, then we have also maintained the relationship between the contributors within a sample, independently of the type of sample normalisation used. This needs to be investigated.

Thus, the aims of this manuscript are to: (1) update the MPS coverage model of Vilsen et al. [1], (2) investigate the use of a discrete compared to continuous distribution, as the amount of input DNA decreases, (3) investigate the effect of sample normalisation on the relationship between the major and minor contributor in two person DNA mixture samples, and (4) investigate the relative difference between the estimated and true mixture proportions in two person mixtures, as the true mixture relationship changed from 1000:1 to 1:1 between the major and minor contributors.

2 Material and Methods

2.1 Experimental data

DNA libraries were build using the ForenSeq™ DNA Signature Prep Kit (Illumina®) Primer Mix A and B. Primer Mix A amplifies markers for human identification (HID), while Primer Mix B amplifies the same HID markers plus ancestry informative markers (AIMs) and markers associated with eye and hair colour. DNA sequencing was performed with the MiSeq FGx (Illumina®) as previously described [17, 18].

DNA was extracted from blood samples and buccal swabs collected on FTA cards from 363 individuals. The samples were amplified and sequenced in duplicate. The data consists of two parts:

- Dilution series of DNA were created from four contributors. The DNA was amplified and sequenced in triplicate using Primer Mix A. The amounts of DNA in each series were 1 ng, 500 pg, 250 pg, 125 pg, 62.5 pg, 31.25 pg, 15.63 pg, and 7.86 pg. A consensus DNA profile from each individual was generated based on all experiments with 1 ng input.

- Fifteen two person DNA mixture samples (in proportions: 1:1,000, 1:100, 1:50, 1:25, 1:12, 1:6, 1:3, 1:1, 3:1, 6:1, 12:1, 25:1, 50:1, 100:1, and 1,000:1) were made with a male and a female contributor.

The total amount of DNA was 1 ng in all cases. The DNA mixtures were amplified and sequenced in duplicate using Primer Mix B. The profiles of the two contributors were known.

Sequences were identified using STRait Razor v3.0 [19] and analysed through the statistical software `R` [20] using the package `STRMPS` found on cran. STRait Razor was only used to identify the STR regions by the flanking region sequences [5] and aggregate the unique sequences. `STRMPS` was used to reduce the number of unique strings, by applying the quality reduction method seen in Appendix A, and to analyse the reduced samples.

2.2 The MPS coverage model

The MPS coverage model includes three major modifications of its CE counterparts: (1) accounting for marker imbalances due to the immaturity of the MPS process, (2) prediction of stutter rates due to the added resolution of the MPS process, and (3) a change in distribution due to the inherent nature of the MPS process.

2.2.1 Marker imbalances

Fig. E.1 shows the coverages of all autosomal markers in the ForenSeq^{TM} DNA Signature Prep Kit for both Primer Mix A and B of 366 samples of 1 ng sequenced in duplicate (not shown here) [18]. The ordinate axis of the figure is shown on a $\log_{10}$ scale and, thus, we clearly observe large difference in coverage between markers. Furthermore, the figure shows that the variation between Primer Mix A and B was consistent across markers (with the exception of markers D22S1045, D5S818, and D9S1122).

We account for the marker imbalances, in accordance with Vilsen et at. [1], by introducing a parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_M)^T$, where M is the number of markers. Furthermore, Vilsen et al. [1] concluded that it would be most beneficial to estimate the marker imbalance, on a workflow database, a database of samples sequenced with the same technology, primer mix, and settings (e.g. the same number of PCR cycles, preparation, and so on), as the sample to be analysed. We denote the maximum likelihood estimate of the marker imbalance parameters found by using a workflow database by $\hat{\boldsymbol{\beta}}$, where we require

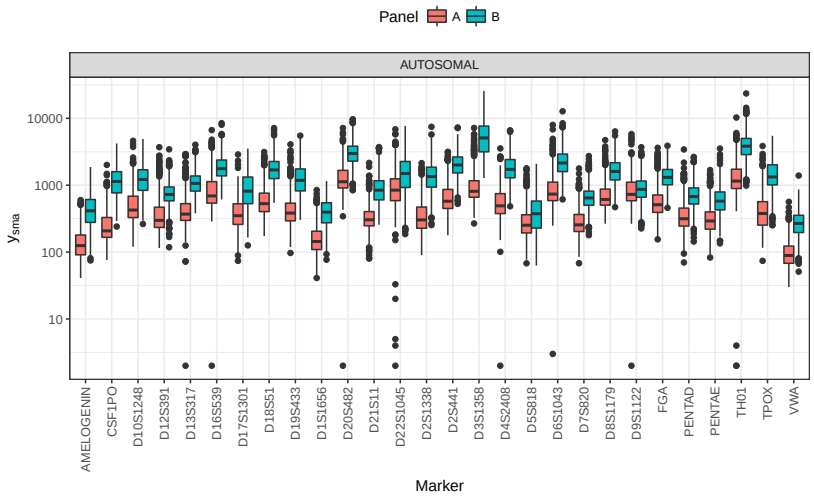

Fig. E.1: The observed coverage for all autosomal markers in both Primer Mix A and B.

$\sum_{m=1}^{M} \hat{\beta}_m = M$ to avoid over-specification of the model.

2.2.2 Stuttering

Due to the added resolution of the MPS process it is now possible to identify multiple stutter sequences of an allele and multiple parental alleles of a stutter sequence. We account for this effect, by introducing a more refined predictor of the stutter ratio, called the Block Length of the Missing Motif (BLMM) [13]. We align a sequence to a potential parent sequence identifying the region (down to a block) and the missing motif. The length of the block, which is missing a motif is the BLMM.

To illustrate the concept, assume we have observed a parental allele sequence:

$$[\texttt{AATG}]_{10}[\texttt{ACTT}]_4,$$

and a stutter sequence:

$$[\texttt{AATG}]_{10}[\texttt{ACTT}]_3.$$

Aligning the stutter to the parent left-to-right, we identify a missing ACTT motif from the $[\texttt{ACTT}]_4$ block. Yielding a BLMM of 4. Thus, the BLMM is an extension of the Longest Uninterrupted Stretch (LUS)

concept. However, instead of using the length of the longest stretch, we use the length of the stretch missing a motif, where the LUS in the example above would be 10.

The relationship between the stutter ratio and the BLMM can be shown to be linear through the point $(1, 0)$. The slope parameters of this linear model can be estimated using a workflow database.

A simple consequence of the increased resolution is that we need to account for every stuttering from every potential parent. That is, the coverage received from stuttering is given as:

$$s^{(1)}_{mac} = \sum_{A \in \mathcal{P}(a)} \xi_m(a, A) g_{mAc}, \tag{E.1}$$

where $\xi_m(a, A)$ is the predicted stutter ratio for sequence a and potential parent sequence A on marker m, $\mathcal{P}(a)$ is the set of potential parents of a, and g_{mAc} is the number of allele A that contributor c has on marker m, i.e. $g_{mAc} \in \{0, 1, 2\}$.

2.2.3 Change in distribution

The coverage of a sequence is synonymous with the count of the sequence. Thus, it is natural to model the coverage as count data. In the past, we have modelled the coverage using a Poisson-gamma distribution of order 2, PG2. The following sections highlight the differences between our previous and present work.

2.2.4 The original coverage model

In summary: The model, presented in Vilsen et al. [1], assumed the following: The coverage of allele a of marker m, denoted y_{ma}, followed a PG2 distribution with mean μ_{ma} and over-dispersion η, where the mean is given as:

$$\mu_{ma} = \nu \hat{\beta}_m \sum_{c=1}^{C} \left[g_{mac} + s^{(1)}_{mac} \right] \varphi_c,$$

where ν can be interpreted as the average coverage of heterozygotes in the sample if the sample contains a single contributor and no artifact (e.g. no stutter sequences, no drop-ins, etc.), φ_c is the relative contribution to the DNA mixture of contributor c, and $s^{(1)}(a, c)$ is given by Eq. (E.1).

2.3 Updating the coverage model

During preliminary analyses and related work, we have found some features of the MPS coverage model lacking in each of the three items recapped above. Therefore, we re-evaluated each of the three items in turn, updating the MPS coverage model.

2.3.1 Marker imbalances

Even though the marker imbalances estimated on a workflow database worked well when comparing the observed and expected Brier scores [1], a closer examination of the marker imbalances exhibited in the DNA mixture samples showed a large deviation from the marker imbalances estimated on a workflow database.

Therefore, we chose to estimate the marker imbalances using the following method-of-moments (MoM) estimate that is only dependent on the sample itself:

$$\beta_m^{\text{MoM}} = \frac{\sum\limits_a y_{ma}}{\frac{1}{M}\sum\limits_{m,a} y_{ma}}.$$

Furthermore, to utilise the information of the database, we created a convex combination of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{\text{MoM}}$:

$$\tilde{\boldsymbol{\beta}}_\lambda = \lambda\hat{\boldsymbol{\beta}} + (1-\lambda)\boldsymbol{\beta}^{\text{MoM}}, \quad \text{where} \quad \lambda \in [0;1] \tag{E.2}$$

and λ represents the weight/belief we assign the workflow database. The closer λ gets to 1, the more weight we put on the workflow database based estimates, $\hat{\boldsymbol{\beta}}$. Whereas the closer it gets to 0, the more emphasis we put on the moment based estimates, $\boldsymbol{\beta}^{\text{MoM}}$. In this paper, the λ parameter is set to 0.2. This choice was heuristic, as estimation should always lead to $\lambda = 0$.

2.3.2 Stuttering

Because of the very skewed mixture relationships analysed in this manuscript, it can be helpful to not only account for stutters but also double stutters. Using the notation in Eq.(E.1), we account for double stutters as:

$$s_{mac}^{(2)} = \sum_{A\in\mathcal{P}(a)} \xi_m(a,A)\left(g_{mAc} + s_{mAc}^{(1)}\right)$$

$$= \sum_{A \in \mathcal{P}(a)} \xi_m(a, A) \left(g_{mAc} + \sum_{B \in \mathcal{P}(A)} \xi_m(A, B) g_{mBc} \right),$$

where it is assumed that the distance between a and A, and A and B, are exactly one motif apart.

Note that this recursion can be extended to any level of stuttering. Assuming that $s^{(0)} = 0$, i.e. implying that the allele itself was not also its own stutter, we can account for $i > 0$ levels of stuttering as:

$$s^{(i)}_{mac} = \sum_{A \in \mathcal{P}(a)} \xi_m(a, A) \left(g_{mAc} + s^{(i-1)}_{mAc} \right).$$

For most practical applications $i \leq 3$, as any contributions from the fourth level would be minuscule. To see how small, assume that we have two sequences four motifs apart, each with a coverage of 1,000 and that we have a stutter ratio of 0.15 in all four levels of stuttering (which is a high stutter ratio), then coverage received from this fourth level sequence would be close to 0.5, compared to a coverage of 1,000.

2.3.3 Change in distribution

Preliminary analyses have shown that using the PG2 distribution will make deconvolution of the profiles in a DNA mixture difficult. Using the PG2 distribution, we saw poorer performance when aggressively optimising with respect to the minor profile at the cost of major. This is due to the very large variance of the PG2 distribution, primarily affecting the major component by reducing its influence on the optimisation. Therefore, we will change from the PG2 distribution to the Poisson-gamma distribution of order 1, PG1. This made the variance of the updated MPS coverage model equivalent to the CE based mixture models.

2.3.4 The updated coverage model

In summary, we assumed that the coverage of allele a of marker m, denoted y_{ma}, followed a PG1 distribution with mean μ_{ma} and over-dispersion γ, where the mean was changed to account for the challenges outlined in Sections 2.3.1 and 2.3.2, yielding:

$$\mu_{ma} = \nu \tilde{\beta}_{\lambda,m} \sum_{c=1}^{C} \left[g_{mac} + s^{(2)}_{mac} \right] \varphi_c, \tag{E.3}$$

where the parameters of the updated coverage model should be interpreted described above.

3 Results

3.1 Between sample normalisation

Vilsen et al. (2017) [2] showed that the coverage of the samples quantified by the Ion PGM HID STR 10-plex were normalised between samples based on the IonExpress Barcode Adaptors. That is, it utilised the first type of sample normalisation discussed in the introduction. This was not the case for the ForenSeq^{TM} DNA Signature Prep Kit.

Fig. E.2 shows the total coverage of each sample in the dilution series experiments. The figure shows that the total coverage decreased as the amount of template DNA decreased. Futhermore, the figure also shows that as the amount of template DNA increases the total coverage tends towards an upper limit. That is, the samples quantified using the ForenSeq kits utilised the second type of sample normalisation. In this case, the sequenced samples were limited to a maximum of 1 ng of template DNA, per the Illumina Forenseq Prep-kit recommendation [21].

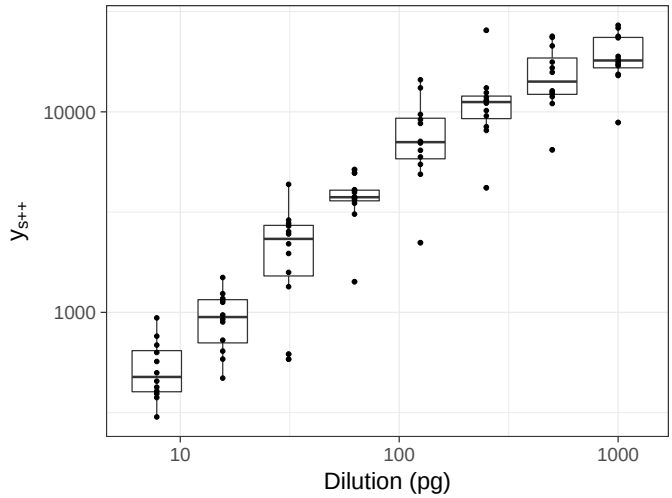

Fig. E.2: The total coverage against the amount of template DNA for all samples in the dilution series experiments.

3.2 Within sample normalisation

We examined the within sample normalisation by comparing the coverage of the alleles of the minor and major contributors, which they did not share and were not in stutter positions. The coverage was scaled by the number of alleles of the contributor and scaled by a marker imbalance parameter, making the observations comparable across markers whether or not the contributor is a hetero- or homozygote.

We compared the scaled coverage with the relative amount of input DNA for each of the contributors to the mixture (Fig. E.3). The figure shows that the relationship between the major and minor is maintained. Thus, we did not see any coverage normalisation between the alleles within the DNA mixture samples.

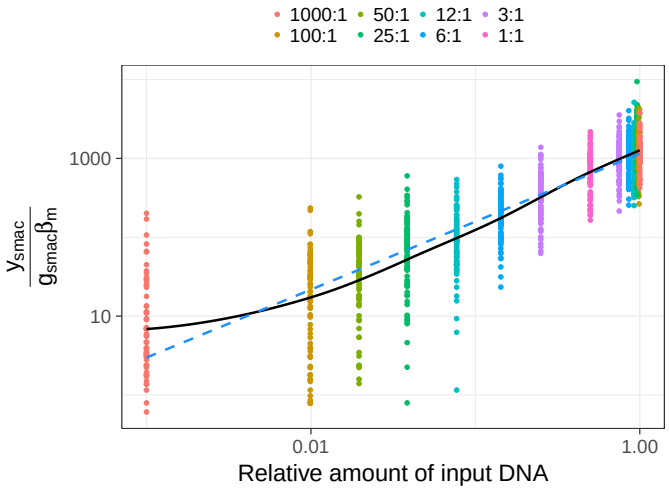

Fig. E.3: The scaled coverage against the relative amount of template DNA.

3.3 Comparing choice of distribution

We wanted to examine the effect of using the gamma distribution as an approximation to the Poisson-gamma distribution. Therefore, we estimated the parameters of the model with the mean described in Section 2.3.4 assuming the coverage followed a gamma and PG1 distribution, respectively. We then constructed and visually inspected P-P plots.

In general, if the amount of template DNA was larger than 31.25 pg, then the two methods were almost identical. This is to be expected

as with large enough coverage the gamma distribution is a good approximation of the PG1 distribution. Fig. E.4 shows examples of P-P plots of the two models. The amount of template DNA was 62.5 pg and 31.25 pg as shown in the top and bottom rows of the figure, respectively. The top row shows that the two models were almost identical (as expected). While in the bottom row we see that the P-P plot of the gamma model is very far from the one-to-one line. This diversion is caused by a dramatic increase in the number of drop-outs in the data. This highlights one of the strengths of using the PG1 distribution, namely that drop-outs can be easily handled by setting the coverage of a drop-out to zero.

3.4 Examining the estimated mixture proportions

We estimated the parameters under the coverage model described in Section 2.3.4. The estimated mixture proportions shown in Fig. E.5 against the true mixture proportions of the major contributor for each of the DNA mixture samples. We see that the coverage model tends to underestimate the true mixture proportions with the exception of the 1:1 mixtures. However, on average the relative difference between the true and estimated mixture proportions was less than 2% (data not shown here).

4 Summary

We updated the coverage model from using the PG2 distribution to using the PG1 distribution, thereby reducing the variance of the coverage model making it comparable with the CE based mixture models. Furthermore, we updated the levels of stutter recursion included to account for double stutters and the method used for estimation of the marker imbalance parameters making them more sample specific. Furthermore, the updated coverage model estimated the mixture proportions to be within a 2% relative difference on average.

We saw that the samples sequenced using the ForenSeq kit were normalised by imposing an upper limit to the amount of template DNA used for each sample. A direct consequence being that if the samples analysed contain less template DNA than upper limit, then the total coverage will be proportional to the amount template DNA.

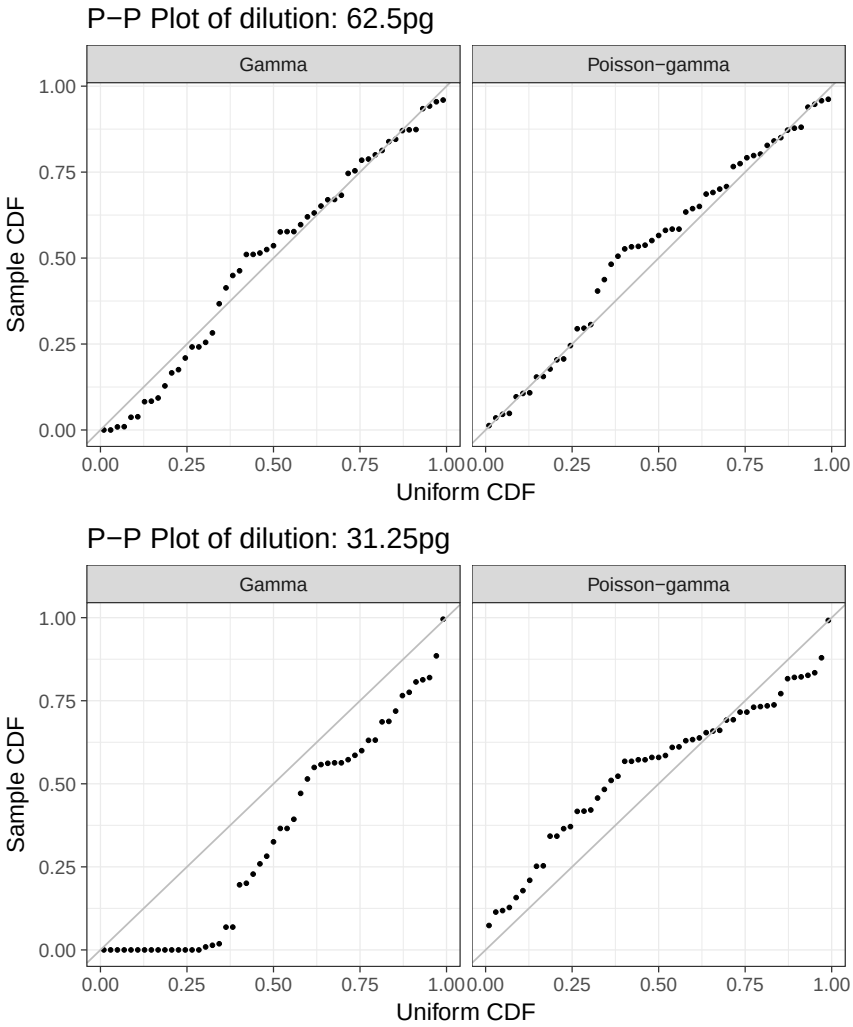

Fig. E.4: P-P plots of the fitted gamma and Poisson-gamma models of 62.5 pg and 31.25 pg DNA.

Note this was not the case for the samples quantified by the Ion PGM [2]. Furthermore, we did not see any within sample normalisation of the samples sequenced by the ForenSeq kit, i.e. the coverage relationship between the major and minor contributors in the analysed two person mixtures were maintained.

Lastly, we showed that if the amount of template DNA is larger than 62.5 pg, then the gamma distribution was a good approximation

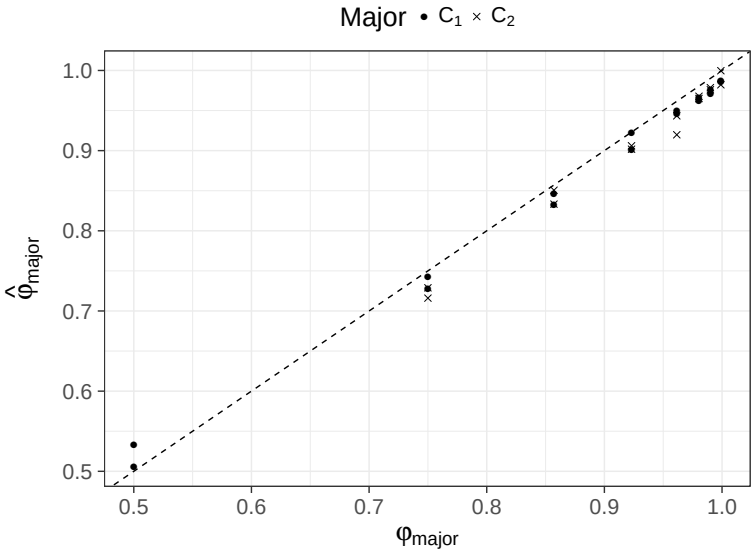

Fig. E.5: The estimated, $\hat{\varphi}_{\text{major}}$, against the true, φ_{major}, mixture proportions of the major contributor. Each point (dot or cross) corresponds to a sample containing DNA from the same two contributors, denoted C_1 and C_2. When C_1 was the major contributor the sample is denoted by a dot. While the sample was denote by a cross when C_2 was the major contributor. The dashed line corresponds to a one-to-one relationship.

of the PG1 distribution yielding almost identical P-P plots. That is, as along as the coverage is large enough it can be modelled using a continuous distribution, in this case, a coverage larger than 25.

Conflict of interest

The authors report no conflict of interest.

Appendix

A Reducing the number of base-calling errors

Our aim was to avoid thresholding by modelling the noise separately from the allele signals and the systematic errors (primarily stutters). Identifying the STR regions of e.g. a ForenSeq-kits `fastq` file with approximately 300,000 to 400,000 reads by searching for flanking regions

(by using STRaitRazor version 3 [19]) results in approximately 4,000 to 6,000 unique strings. Restricting ourselves to the autosomal markers and assuming a single contributor, then approximately one hundred of these strings would be either alleles or systematic errors (leaving approximately 3,900 to 5,900 unique strings contributed to general noise). An important question is: can the number of remaining strings (i.e. the noise) be reduced? That is, could we determine from which '*true*' strings the noise originated? And could we use this information to increase the coverage of the '*true*' strings using the coverage of the strings found among the sequencing errors?

The reduction in the number of unique strings will be based on the base qualities provided in the `fastq` files by analysing the probability of bases having been called erroneously. In the remainder of this appendix, we will for simplicity restrict the discussion to a single marker and strings of equal length. As the aim is to distinguish isoalleles from base calling errors.

A.1 The quality of a base

The quality of a sequenced/called base is defined as a transformation of the probability of the base having been called erroneously. Today, the most common definition of the quality is the Phred score. Given the probability that base n was called erroneously, p_n, the Phred score is defined as:

$$q_n = -10 \log_{10}(p_n). \tag{E.4}$$

Thus, given quality, q_n, the probability, p_n, is easily recovered:

$$p_n = 10^{-q_n/10}. \tag{E.5}$$

Given a string $\boldsymbol{s}$ (seen as a vector of characters i.e. bases) of length N and the corresponding vector of probabilities $\boldsymbol{p}$, we want to find the probability that every base of $\boldsymbol{s}$ was called correctly. If the bases could be considered independent, this probability would simply be:

$$\mathbb{P}\left(\boldsymbol{s} \text{ is correct}\right) = \prod_{n=1}^{N} (1 - p_n). \tag{E.6}$$

However, if a base is called erroneously, it may not necessarily be reflected in the quality of that base, but it may affect the quality of

the surrounding bases due to a sliding window. Thus, the base qualities are not independent. Therefore, we will instead of using p_n directly compensate for the quality of the surrounding bases by using π_n, which is defined as:

$$\pi_n = \max_{j \in \bar{\delta}(n,l)} \left\{ \frac{p_j}{|n-j|+1} \right\},$$

where $\bar{\delta}(n,l)$ is a neighbourhood of n, including n itself, of size at most $2l$ bases. That is, we weight the probabilities in a neighbourhood around base n by its distance to n and take the largest weighted probability as π_n.

Given two strings $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, we want to calculate the probability of $\boldsymbol{s}_i$ actually being the string $\boldsymbol{s}_j$ with some number of miscalled bases denoted $\boldsymbol{s}_i \equiv \boldsymbol{s}_j$:

$$\mathbb{P}\left(\boldsymbol{s}_i \equiv \boldsymbol{s}_j \text{ and } \boldsymbol{s}_j \text{ is correct}\right) = \mathbb{P}\left(\boldsymbol{s}_i \equiv \boldsymbol{s}_j | \boldsymbol{s}_j \text{ is correct}\right) \mathbb{P}\left(\boldsymbol{s}_j \text{ is correct}\right). \tag{E.7}$$

Assuming we know the probability of error of each base of the string $\boldsymbol{s}_i$, denoted $\boldsymbol{p}_i$, then the conditional probability in Eq. (E.7) can be written as:

$$\mathbb{P}\left(\boldsymbol{s}_i \equiv \boldsymbol{s}_j | \boldsymbol{p}_i, \boldsymbol{s}_j\right) = \prod_{n=1}^{N} (1-\pi_{in})^{1-I_n} (\pi_{in})^{I_n}, \tag{E.8}$$

where $I_n = \mathbb{I}\left[s_{in} \neq s_{jn}\right]$, i.e. a function indicating whether s_{in} is equal to the '*truth*' s_{jn}.

The probability that the string '*is correct*' could be based entirely on the quality, in which case it would be given by Eq. (E.6). However, defining it in this way would entirely ignore the information found in the coverage. Therefore, we have defined the probability of a string being correct as:

$$\mathbb{P}\left(\boldsymbol{s}_j \text{ is correct}\right) = w_j \mathbb{P}\left(\boldsymbol{s}_j \text{ does not contain base errors}\right), \tag{E.9}$$

where $\mathbb{P}\left(\boldsymbol{s}_j \text{ does not contain base errors}\right)$ is given directly by Eq. (E.6) and the weight, w_j, of $\boldsymbol{s}_j$ is given by:

$$w_j = \frac{y_j}{\sum_{k=1}^{K} y_k}, \tag{E.10}$$

where y_k is the coverage of string $\boldsymbol{s}_k$ and K is the number of strings of length N. This further emphasises that our approach is heuristic and not analytic.

A.2 Reduction approach

At a given STR marker, assume we have observed a set of strings of length N, $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_K)$ and the coverage of each string given by $\boldsymbol{y} = (y_1, y_2, ..., y_K)$. A matrix of joint probabilities is constructed as given by Eq. (E.7) denoted $V \in \mathbb{R}^{K \times K}$. Given the matrix V, we find the index of largest probability of every column, i.e.

$$k(i) = \arg\max_j \{v_{ji}\}.$$

As the largest probability of the i'th column is $k(i)$, we say that the string $\boldsymbol{s}_i$ is actually the string $\boldsymbol{s}_{k(i)}$, but called with errors, and we remove $\boldsymbol{s}_i$ from further consideration. In cases where the coverage is low, it can be beneficial to retain the information of the coverage of the string $\boldsymbol{s}_i$ and add the coverage of $\boldsymbol{s}_i$ to the coverage of $\boldsymbol{s}_{k(i)}$. The probability matrix V can be constructed by calculating every pairwise combination of Eq. (E.7) for the strings in $\boldsymbol{S}$.

However, we are generally not interested in finding new variants when analysing mixtures. Therefore, we propose the following scheme using a database of '*true*' (or '*trusted*') STR variants.

Assuming we have database of '*true*' STR variants (of the specified marker and length) denoted $\mathcal{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_M)$. At this point, we can take the problem in two directions dependent on whether we want to allow for the survival of any variants. If we do not want this, we just want to calculate the probability of every string in the set $\boldsymbol{S}$ being a variant of a string in the set $\boldsymbol{S} \cap \mathcal{T}$. If we want to allow for the survival of variant strings, we still need the first part, but in addition, we also want the probability of the strings not in $\boldsymbol{S} \cap \mathcal{T}$ (i.e. $\boldsymbol{S} \backslash \mathcal{T}$) being called correctly. Note: if $\boldsymbol{S} \cap \mathcal{T} = \emptyset$, then we would have to calculate every possible pairwise combination.

By the definition of $\mathcal{T}$, it follows that $\mathbb{P}(\boldsymbol{s}_i \text{ is correct}) = 1$ for all $\boldsymbol{s}_i \in \boldsymbol{S} \cap \mathcal{T}$. Thus, calculating the joint probability in Eq. (E.7) is reduced to calculating the conditional probability in Eq. (E.8). That is, we calculate $\mathbb{P}(\boldsymbol{s}_i \equiv \boldsymbol{s}_j | \boldsymbol{p}_i, \boldsymbol{s}_j)$ for every $\boldsymbol{s}_i \in \boldsymbol{S}$ and $\boldsymbol{s}_j \in \boldsymbol{S} \cap \mathcal{T}$, creating a matrix, $V^{(1)}$, of size $I \times N$, where I is the number of elements

in $\boldsymbol{S} \cap \mathcal{T}$ (i.e. $I = |\boldsymbol{S} \cap \mathcal{T}|$). Assuming, without loss of generality, that the matrix (and the set $\boldsymbol{S}$) is ordered such that the first I columns (elements) corresponds to the strings in $\boldsymbol{S} \cap \mathcal{T}$, then the matrix will have the following structure:

$$V^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 & \mathbb{P}\left(\boldsymbol{s}_{I+1} \equiv \boldsymbol{s}_1 | \boldsymbol{p}_{I+1}, \boldsymbol{s}_1\right) & \mathbb{P}\left(\boldsymbol{s}_{I+2} \equiv \boldsymbol{s}_1 | \boldsymbol{p}_{I+2}, \boldsymbol{s}_1\right) & \cdots & \mathbb{P}\left(\boldsymbol{s}_{N-I} \equiv \boldsymbol{s}_1 | \boldsymbol{p}_{N-I}, \boldsymbol{s}_1\right) \\ 0 & 1 & \cdots & 0 & \mathbb{P}\left(\boldsymbol{s}_{I+1} \equiv \boldsymbol{s}_2 | \boldsymbol{p}_{I+1}, \boldsymbol{s}_2\right) & \mathbb{P}\left(\boldsymbol{s}_{I+2} \equiv \boldsymbol{s}_2 | \boldsymbol{p}_{I+2}, \boldsymbol{s}_2\right) & \cdots & \mathbb{P}\left(\boldsymbol{s}_{N-I} \equiv \boldsymbol{s}_2 | \boldsymbol{p}_{N-I}, \boldsymbol{s}_2\right) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \mathbb{P}\left(\boldsymbol{s}_{I+1} \equiv \boldsymbol{s}_I | \boldsymbol{p}_{I+1}, \boldsymbol{s}_I\right) & \mathbb{P}\left(\boldsymbol{s}_{I+2} \equiv \boldsymbol{s}_I | \boldsymbol{p}_{I+2}, \boldsymbol{s}_I\right) & \cdots & \mathbb{P}\left(\boldsymbol{s}_{N-I} \equiv \boldsymbol{s}_I | \boldsymbol{p}_{N-I}, \boldsymbol{s}_I\right) \end{bmatrix}.$$

If we do not want to allow for any variants, then we set $V = V^{(1)}$ and find the '*true*' string for every string in $\boldsymbol{S}$ as described above.

If we want to allow the variant strings (i.e. strings in $\boldsymbol{S} \backslash \mathcal{T}$) a chance to survive, we need the probabilities of these strings having been called correctly. Note we are not interested in the probability of the strings in $\boldsymbol{S} \cap \mathcal{T}$ are called correctly, as they are assumed correct given $\mathcal{T}$. Furthermore, we were not interested in the probability of a variant string was in fact just another variant string having miscalled bases.

Thus, for each string in $\boldsymbol{S} \backslash \mathcal{T}$, we calculate the probability of the string was called correctly, using Eq. (E.9). Assuming, as before, that the first I columns corresponds to the strings in $\boldsymbol{S} \cap \mathcal{T}$, we have defined a matrix $V^{(2)}$ of size $(K - I) \times K$, taking the form $V^{(2)} = [O \; J]$, where O is a $I \times I$ matrix of zeroes and J is the diagonal matrix $\{\text{diag}(\mathbb{P}\left(\boldsymbol{s}_j \text{ is correct}\right)) \mid j \in \boldsymbol{S} \backslash \mathcal{T}\}$. The matrix V is then constructed as:

$$V = \begin{bmatrix} V^{(1)} \\ V^{(2)} \end{bmatrix},$$

and the index of the largest probability is found as discussed above.

Using this approach, we can reduce the number of unique strings from 4,000-6,000 to less than 500. At the same time, we keep the coverage of the sample the same by adding the coverage of the strings in $\boldsymbol{S} \backslash \mathcal{T}$ having been assigned a '*true*' string in $\boldsymbol{S} \cap \mathcal{T}$ to the '*true*' string.

References

[1] S. Vilsen, T. Tvedebrink, P. S. Eriksen, C. Hussing, C. B. rsting, and N. Morling, "Modelling allelic drop-outs in STR sequencing data generated by MPS," *Forensic Science International: Genetics*, 2018.

[2] S. Vilsen, T. Tvedebrink, H. Mogensen, and N. Morling, "Statistical Modelling of Ion PGM HID STR 10-plex MPS Data," *Forensic Science International: Genetics*, 2017.

[3] K. Lazaruk, P. S. Walsh, F. Oaks, D. Gilbert, B. B. Rosenblum, S. Menchen, D. Scheibler, H. M. Wenz, C. Holt, and J. Wallin, "Genotyping of forensic short tandem repeat (str) systems based on sizing precision in a capillary electrophoresis instrument," *Electrophoresis*, vol. 19, no. 1, pp. 86–93, 1998. [Online]. Available: http://dx.doi.org/10.1002/elps.1150190116

[4] J. Butler, *Fundamentals of Forensic DNA Typing*. Academic Press, 2009.

[5] S. Fordyce, M. Avila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F. Petersen, E. Willerslev, A. Hansen, N. Morling, and M. Gilbert, "High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform," *BioTechniques*, vol. 51, pp. 127 – 133, 2011.

[6] D. M. Bornman *et al.*, "Short-read, high-throughput sequencing technology for STR genotyping." *BioTechniques*, pp. 1–6, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22668513

[7] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, and D. Deforce, "Forensic STR analysis using massive parallel sequencing," *Forensic Science International: Genetics*, vol. 6, no. 6, pp. 810–818, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2012.03.004

[8] E. Rockenbauer, S. Hansen, M. Mikkelsen, C. Børsting, and N. Morling, "Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing," *Forensic Science International: Genetics*, vol. 8, no. 1, pp. 68–72, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2013.06.011

[9] S. Dalsgaard, E. Rockenbauer, A. Buchard, H. S. Mogensen, R. Frank-Hansen, C. Børsting, and N. Morling, "Non-uniform phenotyping of D12S391 resolved by second generation

sequencing," *Forensic Science International: Genetics*, vol. 8, no. 1, pp. 195–199, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2013.09.008

[10] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Børsting, and N. Morling, "Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles," *Forensic Science International: Genetics*, vol. 12, pp. 38–41, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2014.04.016

[11] M. Scheible, O. Loreille, R. Just, and J. Irwin, "Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers," *Forensic Science International: Genetics*, vol. 12, pp. 107–119, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fsigen.2014.04.010

[12] C. Børsting and N. Morling, "Next generation sequencing and its applications in forensic genetics," *Forensic Science International: Genetics*, vol. 18, pp. 78 – 89, 2015.

[13] S. Vilsen, T. Tvedebrink, P. S. Eriksen, C. Hussing, C. B. rsting, H. S. Mogensen, and N. Morling, "Stutter analysis of complex STR MPS data," *Forensic Science International: Genetics*, 2018.

[14] R. Cowell, S. Lauritzen, and J. Mortera, "Probabilistic expert systems for handling artifacts in complex dna mixtures," *Forensic Science International: Genetics*, vol. 5, pp. 202–209, 2011.

[15] R. Cowell, T. Graversen, S. Lauritzen, and J. Mortera, "Analysis of Forensic DNA Mixtures with Artefacts," *Royal Statistical Society. Journal Series C: Applied Statistics*, vol. 64, pp. 1 – 32, 2015.

[16] T. Graversen and S. Lauritzen, "Computational aspects of DNA mixture analysis," *Statistics and Computing*, vol. 25, pp. 527–541, 2015.

[17] C. Hussing, C. Huber, R. Bytyci, H. Mogensen, N. Morling, and C. Børsting, "Sequencing of 231 forensic genetic markers using the Illumina® ForeSeq™ workflow - an evaluation of the assay and software," *Forensic Sci. Res.*, 2018, (in press).

[18] C. Hussing, C. Huber, R. Bytyci, , N. Morling, and C. Børsting, "The Danish STR sequence database: Duplicate typing of 363 Danes with the ForenSeq^{TM} DNA Signature Prep Kit," *Int. J. Legal Med.*, 2018, (in press).

[19] A. E. Woerner, J. L. King, and B. Budowle, "Fast STR allele identification with STRait Razor 3.0," *Forensic Science International: Genetics*, vol. 30, pp. 18–23, 2017.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[21] "Illumina ForenSeq DNA Signature Prep Kit," (Accessed: 2018-07-14). [Online]. Available: https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/forenseq-dna-signature.html

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-328-0

AALBORG UNIVERSITY PRESS